AUTOMATIC INTERPRETATION OF NUCLEOSIDE MASS SPECTRA

A thesis submitted in partial fulfilment of the
requirements for the degree of
Doctor of Philosophy in Chemistry
in the
University of Canterbury

M.J. Mitchell

University of Canterbury
1979

ACKNOWLEDGEMENTS

I wish to express my gratitude to a number of people who have contributed in various ways to this work. First a colleague, Dr I.J. Doonan, for many stimulating and fruitful discussions during the early part of this research. Secondly Professor R.J. Ferrier of the Victoria University of Wellington for the supply of samples. Thirdly Mr T.J. Brady for some useful comments on the presentation of results. Fourthly the two proof readers, Mr R.R. Sherlock for his dedicated and efficacious efforts on the manuscript, and Miss A.G. Pattie who corrected the typed copy. And finally I would like to thank my supervisor, Dr R.G.A.R. Maclagan, for his guidance and helpful criticism over the years.

CONTENTS

Page

Page

TABLES

FIGURES

Page

Chapter 8

Chapter 9

Chapter 10

SCHEMES

ABSTRACT

Various heuristic and pattern recognition techniques have been applied to a data base of 125 underivatised nucleoside mass spectra to determine certain aspects of structure from an unknown spectrum. A heuristic program has been written encoding nucleoside mass spectral fragmentations in order to determine molecular weight, formula weight of the purine or pyrimidine base part, and, unsuccessfully, base type. The pattern recognition methods of statistical linear discriminant function analysis, learning machine approach, distance from the mean, and k-nearest neighbour classification have been applied to the same data base divided into training and prediction sets. Analyses were conducted to determine numbers of carbon, oxygen, and nitrogen atoms present in the base alone and in the nucleoside as a whole, substitution patterns, and base type. Prediction success for all approaches was typically in the range 76-86%, or in terms of the figure of merit 0.20 - 0.27.

Chapter 1

INTRODUCTION

1.1 Overview

The structure and functions of ribonucleic acids and deoxyribonucleic acids have been the subject of much attention in recent years. Their structural elucidation can be divided into two parts: identification of the component nucleotides and determination of their sequence in the nucleic acid chain. Identification of the component nucleotides can be performed efficaciously by mass spectrometry after conversion to the corresponding nucleosides to overcome their low volatility. As nucleic acid chains typically consist of one hundred or more component nucleotides a large amount of mass spectral information can be generated, particularly from gas chromatographic-mass spectral (GC-MS) separation of the degraded components.

In addition, modified nucleic acid components of often unknown structure occur in a wide variety of natural and biologically active materials. The identities of such modified components cannot generally be determined from mass spectrometry by comparisons with standard spectra. To this end, and bearing in mind the large number of routine determinations which often have to be performed, it would be very desirable to be able to perform the many mass spectral interpretations automatically or semi-automatically. It is not at present possible to completely elucidate a nucleoside structure by computer techniques. Thus this thesis is an investigation of ways to automatically extract significant information from such mass spectra in order to facilitate later manual interpretation. The parallel problem of sequence determination within a nucleic acid chain, despite some preliminary pattern recognition investigations by Wiebers et al. [1-3], has not been touched, in part because of the lack of success of these workers.

Thus the work described here straddles two fields. Superimposed on the mass spectral fragmentation theory of nucleosides are various computer techniques for data analysis. These are of two types. First, in the heuristic approach, the fragmentations of various classes of nucleosides have been encoded into a program in order to deduce from the spectrum of an unknown compound various aspects of its structure. Secondly, by

pattern recognition analysis, certain structural features have been correlated with the mass spectra of a "training" set of molecules and the classification functions so developed tested on a "prediction" set of unknowns.

Low resolution 70 eV electron impact mass spectra of underivatised nucleosides were obtained almost solely from the chemical literature of the period 1962-78. Many published spectra were rejected either on the grounds of excessive structural variation, such as derivatisation, or of too few peaks in the spectrum. Most guanosines, for example, fell into this latter category. A complete list of the 125 compounds whose spectra were used in this work is given in appendix I. This comprises systematic and, where applicable, trivial names, and reference(s) to the published spectra. The elemental compositions of the 125 nucleosides lay in the range $C_{9-32}H_{11-46}O_{2-8}N_{1-8}S_{0-1}F_{0-1}Cl_{0-1}Br_{0-1}$ and the molecular weights in the range 211-755 amu. Other sources of spectra such as data libraries like the MSDC-NIH-EPA collection (subsection 2.3.2) are not as yet available in New Zealand.

A considerable body of research has previously been conducted into the computer interpretation of mass spectra. Three conceptually different approaches have been adopted. Aside from the heuristic and pattern recognition methods used here,there is also the library search technique. This involves the comparison of an unknown spectrum with those in a large data base or library of reference spectra, until a more or less exact match is found. This technique requires a very large data base and has not been used here. Four pattern recognition methods have been adopted in the present work: statistical linear discriminant function analysis, the learning machine approach, distance from the mean, and the k-nearest neighbour classification. The results are in general comparable with similar literature studies although differences in the structural features investigated and in the size and composition of the data bases have often made comparisons difficult.

1.2 Arrangement of the Thesis

Two reviews and the two types of computer technique are covered in the chapters of this thesis. These are:

(a) a review of general computer applications in other types of spectroscopy and chromatography as well as mass spectrometry (chapter 2),

(b) a review of the fragmentations of nucleosides (chapter 3),

(c) description and results of the heuristic programming (chapter 4), and

(d) description and results of the four pattern recognition methods (chapters 6-9).

In addition the methods have been compared in the final chapter, and chapter 5 contains notes on the data base and on the concepts of pattern recognition. The individual chapters will now be outlined more fully.

Chapter 2 is an extensive though by no means exhaustive review of the use of computers in various analytical techniques. The large quantity of data generated by such techniques makes automation imperative, and this chapter seeks to give a broad overview of the computer methods employed and the analytical methods to which they have been applied. To cover as representative a range as possible attention has not been limited to mass spectrometry. Chromatography, IR, NMR, and X-ray fluorescence spectroscopy are the fields in which some of the most significant advances have been made and these are the ones dealt with here. Coverage is restricted to methods of interpretation of the data obtained rather than to data acquisition and control. Over two hundred original papers have been reviewed. More attention is given to those methods used in this work, such as pattern recognition and heuristic programming, but mention is also made of library search techniques, mathematical modelling, and the like.

Chapter 3 is a review of the mass spectral fragmentations of nucleosides, a topic which has been often previously reviewed. The aspects covered in this chapter are those which pertain particularly to the heuristic studies of chapter 4. In addition, the pattern recognition studies of chapters 6-9 utilise many of the structurally relevant mass positions detailed here.

Chapter 4 contains the results of the heuristic programming approach. The type of programming described in this chapter is very time consuming and extension to a larger number of compound classes would be exceptionally tedious. An externally supplied program for molecular weight determination is compared with the specifically written procedures. This chapter contains all the heuristic work as opposed to the pattern recognition studies of chapters 6-9.

Chapter 5 is a collection of matters fundamental to the pattern recognition studies. It encompasses a description of the data base, an

outline of pattern recognition methodology, and an explanation of the evaluation measures used to gauge the success of the pattern recognition studies. It serves furthermore as a delimiter between the heuristic and the pattern recognition approaches, a break to emphasise their fundamental conceptual differences. The data base obviously is the same as that used for the heuristic studies of chapter 4, but significant aspects of it such as the division into training and prediction sets pertain particularly to the pattern recognition analyses.

Chapters 6-9 contain the four pattern recognition approaches, viz. statistical linear discriminant function analysis, learning machine approach, distance from the mean, and the k-nearest neighbour classification, respectively. These were utilised either as standard programs (chapters 6 and 7) or were coded in ALGOL during this present work from established theory (chapters 8 and 9).

Finally, Chapter 10 is a comparison of the methods, not only amongst the four pattern recognition techniques but also between the pattern recognition and the heuristic approaches in so far as this is possible. A general conclusion and evaluation of the results obtained is offered.

Chapter 2

APPLICATIONS OF COMPUTERS IN CHROMATOGRAPHY AND SPECTROSCOPY

2.1 Introduction

The use of computers both for on line control of analytical procedures and for interpretation of the results obtained has greatly increased in recent years. This chapter is a review of a number of fields in which computer techniques have found application. Attention has been largely restricted to the interpretation aspects rather than to data acquisition and control, although some of the more recent and significant advances in these latter areas are referred to.

Coverage is hopefully representative though by no means exhaustive; chromatography, NMR, IR, and X-ray fluorescence spectroscopy have been taken as exemplifying a cross-section of the principal methods now in use for automatic spectral interpretation and analysis. Most of the major approaches to mass spectral interpretation have been covered, and those particularly relevant to the present work are treated in greater depth. These include the deductive programs such as the DENDRAL project (subsection 2.3.6) of Lederberg et al. and Lageot's silicon heterocyclics program (subsection 2.3.7) for mass spectrometry, and a heuristic IR interpretative program developed by Woodruff and Munk (subsection 2.2.3). Also included are some of the many diverse applications of pattern recognition, such as the characterisation of liquid phases in gas chromatography (GC) (subsection 2.2.1), the interpretation of carbon-13 NMR spectra (subsection 2.2.2), and the identification of oil pollutants from their IR spectra (subsection 2.2.3). The many applications of pattern recognition to mass spectrometry are reviewed in subsection 2.3.3.

2.2 General Chromatography and Spectroscopy

2.2.1 Chromatography GC data acquisition by mini-computer, including data sampling and smoothing, peak detection and integration, baseline determination and separation of overlapped peaks, has recently been reviewed by Caesar [4] and others [5]. Computing

integrators, a progression from digital integrators, have been reviewed by Gill and Hettinger [6], and a recent ACS symposium [7] on micro-processors has illustrated their specific functionality, programmatic flexibility, and increasing use for GC analysis.

The general elution problem [8] in column chromatography has two aspects. These are (a) the loss of resolution of those components of a mixture eluting first, and (b) the inordinate amount of time taken by those components of high retention with consequent excessive band broadening due to diffusion. Two common remedies are gradient elution and temperature programming. The former uses a varying mixture of two or more liquid phases, non polar at first, changing to a predominance of the polar phase, with consequent increase in eluting power throughout the run. Programs controlling the mixing of the mobile phases have been described by Scott [9], and Huang and Fagerson [10]. Micro-processor control of temperature gradient has recently been detailed by Dulson [11], and Sibley et al. [12] have presented a simulation of retention times with varying temperature. An alternative approach to the general elution problem is to use a mixture of stationary liquid phases, and for gas-liquid chromatography (GLC) computer optimisation of such a mixture has been carried out by Molera and associates [13].

The excessive and confusing number of liquid phases in GLC has long been a problem [14,15] and various pattern recognition characterisations have been propounded to group them according to retention behaviour. The Kovat retention index [16] of a solute on a liquid phase can be represented [17] by a linear free energy relationship involving solute and liquid phase "polarity factors". Consequently a liquid phase can be characterised by a set of solute retention indices [18,14] and hence by a point in multi-dimensional space. The nature and number of these solutes have been extensively investigated by nearest neighbour similarity measures [19,20], factor analysis [21], principal component analysis [22], and a semi-theoretical classification [23].

2.2.2 NMR A number of reviews of mini-computer NMR data systems have recently been presented [24], and Ernst [25a] and Shaw [25b] have reviewed Fourier Transform (FT) NMR dealing especially with enhancement of the signal to noise ratio and control of the pulsing unit. The basic FT algorithm of Cooley and Tuckey [26] is well established.

Among the more interesting data systems described recently is one for flowing NMR [27], and a fast least squares technique for determination of the spin-lattice relaxation time from pulsed NMR data [28].

A large data library has only recently become available for carbon-13 NMR with the NIH-EPA collection [29,30]. This contained 4024 spectra in 1978 and is accessible by an interactive telephone link in the United States and Europe. Smaller carbon-13 NMR data libraries have also been established by a group of German workers [31] and by Jezl and Dalrymple [32]. A simple use of proton shifts has been made by the Stanford University DENDRAL project as a part of their mono-functional acyclic amine program [33](cf. subsection 2.3.6) and a similar although smaller program has been constructed by a group of Japanese workers [34].

Pattern recognition [35] has been applied primarily to carbon-13 NMR although a factor analysis [36] of the proton spectra of simple alkanes and some mono-functionals has been conducted by Malinowski et al. [37]. Kowalski and Reilly performed an early linear discriminant classification on proton NMR [38]. Wilkins and co-researchers carried on from this latter work, making use of the well known sensitivity of carbon-13 NMR to structural variation. They represented such spectra as points in a 200-dimensional space, for a chemical shift range of 0-200 ppm. For the identification of various functionalities they investigated types of linear discriminant functions i.e. decision hypersurfaces (subsection 5.1.1), and various pre-processing techniques [39]. More sophisticated classification approaches have lately been taken [39e,40].

Accurate parametric formulae have been established by Lindeman and Adams [41] for the carbon shifts of alkanes, on the basis of numbers and types of alkyl substituents. Programs for the identification of alkanes from their carbon-13 NMR spectra based on this parameterisation have been described by Burlingame et al. [42], Suprenant and Reilley [43,44], Sasaki et al. [45], and Djerassi and co-workers [46,47]; all involve generation of structures from a given carbon number. Some simple mono-functionals have also been similarly treated [43,44,46,47] with varying success. A rather more restricted parameterisation of substituted norbornanes has lately been achieved [48].

An interesting modern development is the in vivo and in vitro proton NMR imaging of medical and biological tissues, originated by

Lauterbur in 1973 [49]. An image from a specific region of a sample can be obtained if a magnetic field gradient is applied while the NMR measurement is being made. Two approaches appear the most useful. Hinshaw's sensitive point method [50] defines a small volume by means of three orthogonal gradients, and scans it through the sample. A method involving selective irradiation of the sample and subsequent computer image reconstruction of the FT of the resultant free induction decay has been developed by Mansfield [51]. One potential application of the technique is to the early detection of tumours [49] by variations in the spin-lattice relaxation time, although this has recently been questioned [52].

Computer simulation of proton NMR spectra has been well reviewed by Haigh [53] who discusses solution of the secular equations, determination of the energy levels, and calculation of the spectral frequencies and intensities which are then fitted to the experimental values by a least squares analysis [54]. Several general programs are available [55]. A paper discussing the reliability of such simulations has recently appeared [56], and the approach required when simulating spectra involving lanthanide shift reagents has been described [57]. Spectral simulation involving line shape fitting for resonances broadened by rapid conformational change has been reviewed [58] and standard programs are available [55]. The equations of Gutowsky and Helm [59] for line shapes broadened by chemical exchange [60] have been programmed by Moore [61].

2.2.3 IR Mini-computer interfaces to IR spectrophotometers have been described by Mattson [62], and Isenhour and co-workers have presented [63] an efficient vidicon tube method for direct digitisation of IR spectra. The usual problem with difference spectroscopy of unknown mixtures [64] is the need to access large files of known reference spectra for the necessary subtraction, and several programs eliminating this need have recently been described. One utilises data from partial fractionation of the mixture [65] while another allows fingerprinting by interactive subtraction [66]. FT IR spectroscopy has been variously reviewed [25a, 67] and recent developments include analysis of sub-nanogram samples [68]. Young has described [69] a derivation of gas phase rotational and collisional times from the FT of IR absorption bands, and Isenhour and

de Haseth have devised a method [70] to reconstruct gas chromatograms from single scan GC/IR interferograms using a Gram-Schmidt vector orthogonalisation.

The largest IR data base is the ASTM collection [71] of 92,000 spectra, and other major files have been catalogued by Gevantman [72]. Private libraries have more recently been established by Penski et al. [73] and Fox [74] amongst others. Tanabe and Saeki have published [75] a correlation coefficient matching method for IR spectral retrieval. Isenhour and colleagues have compared [76] two nearest neighbour similarity measures for classification of an unknown according to its best match in a library. A representation of an IR spectrum as, say, a 139-dimension vector for 0.1 μm intervals in the region 2.0 - 15.9 μm renders such spectra amenable to pattern recognition methods, and linear discriminant function [77] and maximum likelihood [78] classifications have been attempted. The former applied particularly to identification of oil pollutants, and the statistical significance of spectral differences between oils from various sources. A comparison of the classification methods has also appeared [79]. In a very different approach to library or pattern recognition techniques, Gray [80] and Woodruff and Munk [81] have reproduced in a program the steps an actual spectroscopist would take to assign structures to IR[81] or IR and NMR [80] spectra. Both programs require as input the empirical formulae, and identify various functional groups directly from their spectral manifestations.

2.2.4 X-Ray Fluorescence Data acquisition and intensity-composition correlation in alloys and geological and biological samples etc comprise the bulk of the computer applications in X-ray fluorescence spectroscopy.

On line analysis and control have been recently reviewed [82] and one of the major problems in the collection of spectral data, deconvolution of overlapped peaks in the presence of high and fluctuating background, appears largely solved [83,84]. Quantitative elemental analysis by X-ray fluorescence depends upon the accurate determination of the non linear relationship [85] between spectral intensities and composition. Two numerical techniques are in standard use, the empirical coefficients [86,87] and the fundamental parameters [88,86] methods. The former expresses the concentration c_i of a given element i, as derived from the

relative intensity R_i of one of its spectral lines, as a linear combination of the concentrations c_j of the other elements present

$$\frac{c_i}{R_i} = \sum_j \alpha_{ij} c_j \qquad (2.2.1)$$

A number of standards, equal at least to the number of elements to be determined, are required, and the coefficients are valid only over a narrow compositional range. To minimise these disadvantages Rasberry and Heinrich [85,89] have explicitly incorporated absorption and secondary fluorescence in an empirical coefficients program.

The fundamental parameters method on the other hand uses theoretically derived equations [88] and requires knowledge of only the secondary fluorescence yield, primary spectral distribution, and matrix absorption coefficients. Only elemental standards are needed and compositional range is unlimited, as accuracy depends only upon the accuracy with which the parameters are known. Model programs have been described [90] and applications include analysis of fused geological samples [91] and metal alloys [92]. The bias arising from inaccurate parameters can be minimised if even one multi-component standard is available [93] and this fact has recently been used by Criss and co-workers [94] in a fundamental parameters program incorporating elements of the empirical coefficients approach. A modified fundamental parameters program has been used [95] for the determination of trace elements in plant materials, by calculating the absorption of each heavy element and attributing the matrix effect of the remainder of the sample mass to cellulose. For this an additional absorption correction was computed. The calculation of matrix effects in biological and environmental samples has been treated [96] by a program using coherent and incoherent scatter peaks to estimate light element (atomic number $\leqslant 13$) content.

2.3 Mass Spectrometry

Mass spectrometry has proved a fruitful field for computer analysis [97-100], particularly for the rapid identification of GC-MS components and for the exhaustive consideration of a large number of structural possibilities for a single compound. Attention in this section is directed

towards the interpretative methods of library search, pattern recognition, heuristic and deductive programming, combination programs such as STIRS, and the sequencing of polypeptides and polysaccharides. Only brief mention is made of the more recent and significant advances in data acquisition, and many traditional computer fields such as molecular orbital calculations [101] and quasi-equilibrium theory [102], which have been applied to the fundamental processes rather than to direct spectral interpretation, are omitted entirely.

2.3.1 Data Acquisition On line mini-computer data collection and presentation systems have recently been reviewed [97] and standard programs are available [103]. Modern developments have centred on GC-MS systems [104-106], the detection of metastables [107], more efficient processing [108], isotope analysis [109-111] and the analysis of mixtures [112]. The mini-computer linked mass spectral system at the University of Canterbury, on which some of the nucleosides of subsection 5.2.1 were recorded, has been described by Wright et al. [113].

2.3.2 Library Search The simplest form of spectrum identification is by exact match with a member of a library data file [97,98,114]. The major data bases currently available are the "Registry of Mass Spectral Data" [115] which in its extended form [116], available on magnetic tape, contained in 1977 the spectra of 30,476 compounds, and the MSDC-NIH-EPA collection [29]. This is a combination of the US NIH file [117,118] and the Aldermaston Mass Spectrometry Data Centre (MSDC) magnetic tape collection [119,120]. The Mass Spectral Search System (MSSS) of MSDC-NIH-EPA is accessible by an interactive telephone link in Europe and the United States. McLafferty and colleagues have combined these files with spectra from other sources to establish in 1978 a set of 41,429 spectra [121]. Other accessible data bases have been catalogued by Gevantman [72]. Several libraries combining GC retention indices and mass spectra for on line GC-MS identification are in use [122,123] and the automatic addition of new spectra has been described [124].

To reduce the identification problem from the extreme case of comparison of every peak in an unknown spectrum with every peak in every member

of the data file, much attention has been devoted to efficient spectral coding and to matching techniques and search algorithms, as will be outlined here. Significant developments in these fields have recently been summarised by Blaisdell [124]. Early library work included investigation of the best form of spectral coding, for example the binary representation [125,126] and the selection of the two largest peaks in each 14 amu mass interval [127]. The former is nowadays the more widespread [128,129]. One important form of coding is the ion series spectrum [130,131] which is a fourteen peak spectrum obtained by summing the intensities at every fourteenth mass position. This representation has been used to characterise the components of GC-MS runs into compound classes [130-134] as described more fully in subsection 2.3.7.

Matching procedures have recently been surveyed by Grotch [135] and Gray [136]. The degree of match of an unknown with a file coded by the six or ten most intense peaks can vary according to the order of comparison and the mass ranges considered [137]. A matching method derived from information theory has been applied [138] to a data file coded according to the eight or twenty-five highest peaks, and information theory has also been used [129] to select those mass positions most discriminatory for comparison and retrieval of spectra. This latter problem has also been approached by a maximum likelihood weighting of mass positions so as to maximise separation of different spectra [128]. The uncommon technique of reverse searching, i.e. comparing a library spectrum to an unknown rather than vice versa, has been utilised [139] for the identification of biological components in GC-MS, where high background and interference makes the selection of mass values corresponding to those of a reference desirable.

2.3.3 Pattern Recognition Pattern recognition studies have been conducted more in mass spectrometry than in any other branch of chemistry [140,35]. A high degree of sophistication is now being achieved [141] and use as an on line tool is forseeable in the near future [142]. Not all areas are equally advanced, of course, and many types of analysis such as those of chapters 6-9 are not as yet suitable for on line use. Given below is a summary of the early work in the field followed by an outline of modern developments including dimensionality reduction, non linear separation of data, factor analysis, comparison

of the sophisticated techniques now in use, and spectral prediction from an input formula. A general introduction to pattern recognition concepts and terminology is given in section 5.1.

Initial applications were to simple problems with linearly separable data, and classifications were obtained between mutually exclusive classes with linear discriminant functions. For example, if it is desired to determine the presence or absence in an unknown of, say, a phenyl group, a training set of spectra of compounds both with and without phenyl groups is represented as points in an n-dimension space, for 1-n mass range. Many techniques are available for determining to which of the two classes the unknown belongs. The simplest is the sum spectra or distance from the mean approach [143] (chapter 8) in which the mean ($\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$ ...) of each class is calculated and the generalised euclidean distance

$$\sqrt{\sum_i (\bar{x}_i - x_i)^2} \qquad (2.3.1)$$

of the unknown (x_1, x_2, x_3 ...) from each mean determines to which set it belongs (i.e. whether it contains or lacks a phenyl group).

Numerous refinements to and variations upon this basic approach are possible, such as normalising the spectra so that each contributes equally to the mean [143] or using binary [125b] or other intensity pre-processed [144] spectra. The technique has also been applied by Rotter et al. [145,146] to the mass spectra of steroids in a study similar to that of chapter 8. They have more recently determined decision surfaces between various classes of steroids by a least squares linear regression analysis [147].

A more sophisticated technique is the linear learning machine approach [148,149,141] (chapter 7) in which a decision surface separating the two classes is iteratively developed on the training set by error correction feedback i.e. by modifying the classifier each time it makes a wrong assignment. This feedback technique is necessary for any kind of decision vector more complex than the simple perpendicular bisector of the means of the classes as in the sum spectra method. Another well used classification technique is the k-nearest neighbour approach [150] (chapter 9) in which an unknown is classified according to the k (generally 1,3,5 or 7) spectra in the hyperspace to which it is closest. In the

limit as the training set becomes large this technique, for k = 1, approaches a library search. A comparison [143] of the efficacies of these various simple methods showed the k-nearest neighbour technique, the most computationally expensive, to be the best. This result has not however been duplicated in the present work (chapter 9). A survey of means of evaluation of the more complex methods now in use has appeared [141]. Multi-category classifications [151] are possible using a combination of decision vectors.

Dimensionality reduction or feature extraction [152] is the selection of those mass positions or combinations of mass positions most useful for a given classification. For a set of compounds of molecular weight up to, say, 350 amu, the problem could be treated in a 350-dimension hyperspace. It is very much more efficient, however, to remove unhelpful components and work in a reduced space. Efficient classifications have been achieved with a simplex method using only eleven weight vectors [142], and with a learning machine approach using eight features (i.e. mass positions) extracted manually from the spectra of a restricted set of phosphonates [153]. Much of the chemical data is linearly non separable [150] and consequently "adaptive digital learning networks" have been used [154] and found to compare favourably with linear discriminant functions used on linearly separable data [155]. The "adaptive digital learning network" technique involves computing joint probabilities of occurrence of randomly chosen subsets of binary pattern elements derived from the mass spectra.

Following a suggestion of Rogers and co-workers [156] feature extraction has recently been treated by factor analysis, a "multivariate statistical technique which simultaneously analyses multiple measurements on many compounds" [157]. The basic assumption is that the data and the categories into which it is divided, e.g. class membership and non membership, are related through the variance. Factor analysis involves the extraction of a set of eigenvectors, i.e. components or factors, from a matrix of intensities at various mass positions for a number of compounds, and their compression to give a minimum dimension representation of the original data. A geometric conceptualisation of this is the construction of an orthogonal set of reference axes in a subspace of the original measurements which contains almost all of the original information. Rotation of these axes into certain alignments can relate the factors to

chemically significant properties, and the original data can be regenerated as a check. The technique is useful in that it compresses and classifies data, and is a possible prelude to pattern recognition because of its ability to determine the minimum number of independent components. Interpretation of these factors is claimed to be able to reveal the fundamental variables underlying mass spectra [158]. Various types of selection, compression, and transformation of the original data and of the factors have been compared by Rozett and Petersson for suitability to mass spectra [158]. As an application of their theory they have derived [157] three factors denoting either characteristic mass positions or characteristic compounds from the spectra of 22 benzenoid isomers of $C_{10}H_{14}$, and have used these for classification of related unknowns [159]. Isenhour and associates have used factor analysis in a preliminary investigation of the linearity of the relationship between mass positions and structural features [160], correlating such functionalities as carbonyl, hydroxyl etc with various m/z values. These researchers have also determined the presence and concentration of various components in a series of related mixtures from GC-MS data [161] using the same technique. Factor analysis has also been used for the sequence analysis of oligodeoxyribonucleotides [162] as is discussed more fully in sub-section 2.3.4.

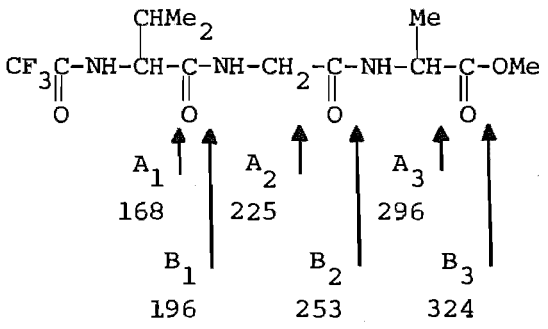
Pattern recognition can be used in the reverse sense to the interpretative procedures outlined above i.e. it can be used to predict mass spectrum from a given input molecular formula [163-166]. This requires the molecular structure to be coded in a vector format suitable for representation in an n-dimensional hyperspace. Decision vectors, one for each mass position, are then trained on a set of such coded formulae and their associated mass spectra to enable the prediction of the presence or absence of a peak at that mass position for a given structure. Approximate intensities can also often be derived. The principal problem is a suitable structure coding technique; that originally chosen [163,164] was termed fragmentation coding and recorded for each molecule such hopefully relevant information as molecular weight, largest ring cycle, aldehyde group presence, etc. Only small molecules (molecular weight $\lesssim$ 250) could be validly treated, and the coding was neither unique nor very accurate, e.g. little information on positions of substituents could be included. Despite these drawbacks spectra were generated with

peaks at 90% of the correct m/z values. Later a modified form of fragmentation coding was introduced [165] which allowed representation of the simultaneous presence of several functionalities, and which utilised feature extraction to reduce training time and enhance predictive ability. These forms of structural coding [163-165] however had had to be carried out manually, and a recent paper [166] has presented an automatic method of deriving substructure codes from a connection table representation.

2.3.4 Sequencing of Biooligomers Sequence analysis by computer of oligopeptides and oligodeoxyribonucleotides from their mass spectra has been approached by two different methods; reconstruction of the chain from spectrometrically determined fragments for the former, and pattern recognition analysis for the latter.

Sequence determination of oligopeptides and proteins, traditionally conducted by stepwise Edman degradations [167,168], has been found amenable to mass spectrometry [169]. This approach has been comprehensively reviewed by Arpino and McLafferty [170, and references cited therein] and others [171,172]. Twin advantages lie in the smaller amounts of material required, compared with the Edman procedure, and, if computer reconstruction methods are used, in the exhaustive consideration of all possible sequences.

Peptides in general undergo two major fragmentation processes, involving cleavage of the amide backbone at one of two possible sites to separate off the C-terminus end. The two resulting series of ions, A and B, are illustrated for a simple oligopeptide in scheme 2.1. Three of the sequencing programs described below [173-175] utilise high resolution data and are based on the fact that



Scheme 2.1: Possible A and B series fragmentations of N-TFA-ValGlyAla-methyl ester. Mass values shown of fragments.

> "with the exception of the isomeric pair leucine and isoleucine, the elemental composition of the side chain identifies each of the common amino acids" [174].

Automatic reconstruction of the chain sequence from such mass spectrometric data has been variously approached by five groups of researchers and these are summarised here. The program of Barber and co-workers [173], starting with a high resolution spectrum of an oligopeptide, identified first the molecular ion and then, given a list of amino acids thought to be present, subtracted each peak in the spectrum in turn from the molecular ion, checking for both A and B ions, until a match with one of the amino acids was obtained. This process was repeated for each determined peak until the full sequence was elucidated. The program was suitable for analysis of cyclic as well as linear peptides, but contained a major drawback in its need to determine the molecular ion. This is often difficult to do unequivocably.

Two groups working independently but in parallel, McLafferty and co-workers [175,176] and Biemann and co-workers [174], advanced an alternative approach beginning at the low mass N-terminus end. Suitable derivatisation enabled the N-terminus ions A_1 and B_1 to be uniquely identified, and each of the twenty possible common amino acids was then added on in turn until a peak was found corresponding to either an A or B series ion. The molecular ion M, or M-CH_3 or M-H_2O in the absence of M, was checked for by adding on the C-terminus group. Because some of the high mass range A and B series ions may not be detectable in sufficient abundance for a high resolution measurement, these were recorded in low resolution and a search made for C-terminus fragment ions to identify that end of the molecule. Fragmentations of the side chain, specific for each amino acid, were also dealt with [174]. The program of McLafferty et al. was later extended to two and three component oligopeptide mixtures [176].

A third, conceptually different, approach has been taken by Nau and Biemann [177,178]. This involved first degrading a protein or polypeptide into small fragments- di- and tri- peptides -by partial acid hydrolysis or enzymatic degradation [179], and derivatising with N-TFA, O-TMS, etc. Secondly, these oligopeptides were analysed by low resolution

GC-MS and each component identified by retention index calculation, comparison of the mass spectrum with that of an authentic sample, etc. Thirdly, the small oligopeptides were joined by a computer program according to the "domino principle" illustrated in scheme 2.2.

```
Phe-Ala-Thr
    Ala-Thr
        Thr-Tyr
        Thr-Tyr-His
            Tyr-His-Try
                etc
```

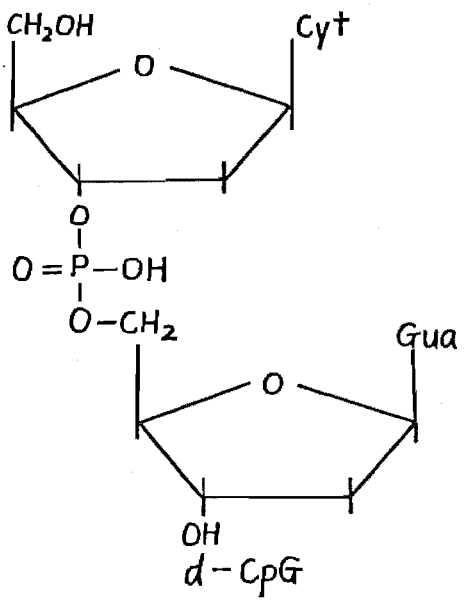
Scheme 2.2: Reconstruction of an amino acid sequence. Small oligopeptides recombined by the "domino principle".

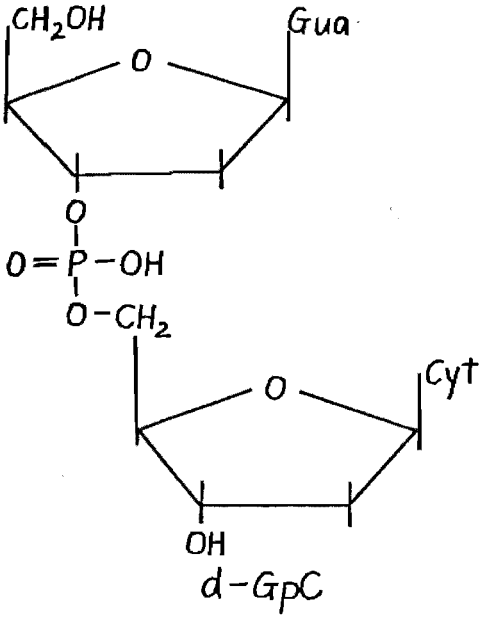
Two advantages of this program lie in its ability to consider all possible linkages, and in its applicability to large, e.g. 50 amino acid, polypeptides and proteins. The initial degradation step is crucial to the success of the method [177], and can be done either randomly, as by partial acid hydrolysis, or by an enzyme such as dipeptidylaminopeptidase I (DAP I)[179] which cleaves the chain into sequential dipeptides. With DAP I the peptide chain is also reduced by one amino acid by means of a single Edman degradation, and parallel degradations on both the reduced and the unreduced chains yield two overlapping series of dipeptides. A similar program had been earlier developed by Dayhoff and Eck [180] and applied in theory only to the structure of the two peptide chains of insulin [181,182]. This involved computer synthesis of all possible oligopeptides arising from the random acid hydrolysis of the insulin chain, and reconstruction of longer and longer chains by overlapping of these fragments.

Amongst the many miscellaneous applications that have been reported is a library file of literature references to peptides which has been stored in and accessed by computer by Weise and Desiderio [183]. This is for the use of mass spectroscopists and others wishing to investigate model compounds, etc.

The enhanced susceptibility of oligodeoxyribonucleotides to mass spectrometric fragmentation as a result of their more labile phosphodiester linkage [184] as compared with oligoribonucleotides, has prompted a pattern

recognition study of their spectra [1-3]. Wiebers and colleagues have reported preliminary investigations into the sequence determination of dideoxyribonucleotides [1] and tri- and tetra-deoxyribonucleotides [3], with mixed results. Ratios of peaks were employed in the sequencing. For example, in the dideoxyribonucleotide case, to distinguish deoxyguanylylcytidine monophosphate (d-GpC) from deoxycytidylylguanosine monophosphate (d-CpG) the peak ratios 162/228, 162/242, 162/271, 191/268, 148/268 and 162/268 were compared with previously determined





thresholds. These and similar compounds of the form d-pXpY and d-pXpYp containing the four common nucleosides adenosine, guanosine, thymidine and cytidine as X and Y served as models for the sequence determination of small oligodeoxyribonucleotides [3] with, however, only partial success. Linear combinations of the features, i.e. peak intensity ratios, determined for the dinucleotides were used for selected oligonucleotides up to $(pApC)_5$. Those combinations of features yielding the clearest distinctions were determined iteratively, considering initially all peaks in the spectrum. The same authors have also conducted a trial factor analysis [158] on a similar set of small oligodeoxyribonucleotides [162] in an attempt to improve prediction, but without success:

"The factor analysis of the normalised-to-sum ions for each nucleoside indicated that any variation in the fragmentation patterns for the nucleosides that were related to the position of the nucleoside in the compound, were less than the experimental variation in the data, and could not be determined by factor analysis" [162].

2.3.5 STIRS The Self-Training Interpretative and Retrieval System (STIRS) developed by McLafferty and colleagues [185-191] for the automatic interpretation of mass spectra, incorporates elements of the heuristic, pattern recognition, and library search approaches [185]. Structurally significant information as described below, selected on the basis of current mass spectral fragmentation knowledge, is extracted from each spectrum of a large library file and stored on magnetic tape. The spectrum of an unknown is also reduced to the same condensed format of structurally significant ions, ion series, and neutral losses, and compared with each member of the data base. If an exact match is found, the compound is considered identified; if not, a number of closely matching spectra are output together with an indication of their degree of similarity to the unknown (the match factor MF), and from structural features common to many of these compounds it is possible to deduce substructural information about the unknown. The program is based upon a general method of manual spectrum interpretation propounded by McLafferty [192]. Formulae of compounds in the library are encoded in Wiswesser Line Notation [193]. Each of the 24,000 spectra at present comprising the data base is condensed by extraction of the following data classes [187]:

(1) Ion series ($\leqslant$ 100 amu) MF1. Included here are each of the fourteen homologous series together with special series such as the so called low and high aromatic sequences at mass values 38, 39, 50, 51, 63, 64, 75, 76 and 39, 40, 51, 52, 65, 66, 77, 78, 79, etc. The five most abundant series, provided they contain at least three peaks each, are ranked according to their intensity sum.

(2) Low mass characteristic ions ($\leqslant$ 89 amu) MF2. These are taken as the three most abundant odd and the three most abundant even mass peaks in this range.

(3) Medium mass characteristic ions (90-149 amu) MF3. The five most abundant peaks of odd and of even mass.

(4) High mass characteristic ions (150 amu - M) MF4. The five most abundant peaks of odd and of even mass.

(5) Small primary neutral losses ($\leqslant$ 65 amu) MF5. The most abundant ions formed by losses from the molecular ion M of o (i.e. M), 1,2,15-64 amu.

(6) Large primary neutral losses (> 65 amu) MF6.

(7) Secondary neutral losses from the most abundant $(M\text{-odd})^+$ ion MF7.

(8) Secondary neutral losses from the most abundant $(M\text{-even})^+$ ion MF8.

(9) Class 8 data of the unknown matched against class 5 data of the reference spectrum MF9.

(10) Fingerprint ions MF10. The most abundant odd and the most abundant even mass ion in each 14 amu interval in the range 90 amu-M.

The program operates in a two pass mode. The first is a library retrieval search using the fingerprint ions, data class 10, only. If a sufficiently high match factor MF10 is found with some library spectrum, the unknown is considered identified. If not, the second pass, designed to indicate structural features of the molecule, commences with the identification [186,192] of the molecular ion. An option is also available to input this manually. Then each of the other data classes above - regarded as "ion series" space, "characteristic ion" space, etc, to adopt the pattern recognition terminology - is scanned for those fifteen reference compounds which are the "nearest neighbours" (cf. chapter 9) of the unknown in each space. If these are tightly clustered in the feature space and are of closely related structure, a high degree of structural similarity to the unknown is indicated. Otherwise no conclusions can be drawn i.e. the absence of specific structural features cannot be determined. If a certain substructural feature, e.g. a phenyl group, amide linkage, etc, is found in a given number of the fifteen best matching compounds, then that feature is indicated with generally a better than 97% probability [185] in the unknown. These analyses are conducted for each data class; an overall match factor calculated from the formula

$$\frac{MF1 + MF2 + 2MF3 + 2\,MF4 + 4MF5 + 2MF6}{12} \qquad (2.3.2)$$

gives generally better prediction than any of the separate data class match factors.

Recent developments undertaken to improve substructure identification [188] include modifications to the characteristic ion [189] and neutral loss [190] data classes, involving variations in the mass ranges, sets of homologous series of neutral losses, etc. A STIRS prediction of the rings plus double bonds value, independent of and complementary to the elemental composition, has also been reported [191] as a possible aid to distinguishing between alternative molecular formulae of equal molecular weight. Evaluation of the performance of the system has been described in theory by McLafferty [194] on the basis of the often contradictory goals of recall (if a substructure is present, how often will this be predicted?) and reliability (if a prediction is made, how often will it be correct?). Furthermore, a comparison of the STIRS system and the k-nearest neighbour pattern recognition technique (chapter 9) for the identification of 500 unknowns containing twenty substructures has been made [195], with the former generally proving superior. This was attributed to the prior selection by STIRS of structurally significant data.

A complementary computer program to STIRS, for the rapid and reliable routine automated identification of given components of relatively complex mixtures without prior separation, has been implemented by the same researchers [196,197,109]. This Probability Based Matching (PBM) system employs statistically determined peak occurrence probabilities [198] to assign a semi-quantitative value to the similarity between an abbreviated spectrum of the compound being searched for, and the unknown. 10-15 peaks, characteristic of the compound and of low occurrence probability, are looked for in a reverse search procedure in the spectrum of either a pure sample or a mixture. The method is based on the principle that if two peaks have statistical occurrence probabilities, as determined from the study of 18000 spectra [198], of $p(1)$ and $p(2)$, the statistical probability of their both being present in an unknown is $p(1)p(2)$. If this value is small and both peaks are found in the unknown spectrum there is a high degree of confidence in the identification obtained. Despite the implicit and partially invalid assumption of independence of mass spectral peaks, the results attained [197] on components of biological mixtures in as low as 10% concentration, and other considerations [185], lend credence to the method, particularly as a relative measure.

McLafferty and co-workers in a related study [121], have defined an algorithm for quantifying the quality of reference mass spectra in a library file. The seven factors of source of spectrum, ionisation conditions, high molecular weight impurities, illogical neutral losses, isotopic abundance accuracy, number of peaks, and the lower mass limit are assigned numerical values and an overall "quality index" calculated for the spectrum. According to the value of this index spectra were included in or rejected from their data base.

2.3.6 Heuristic Work of Lederberg et al. The artificial intelligence project at Stanford University under the direction of Professor J. Lederberg has produced a series of papers [199-201, 33, 202-207, 47, 208-216] describing their programs for the automatic interpretation of mass spectra and related problems. The foundation of their approach is the ability to exhaustively and irredundantly generate all possible molecular or fragment structures from a given formula. Their work can be divided into three broad sections:

(1) high and low resolution mass spectral interpretation based on encoded fragmentation rules for various classes of molecules,

(2) structure generation per se and its direct application to structural elucidation problems, and

(3) automatic construction of fragmentation rules from the mass spectra of series of previously unstudied compounds.

These three sections are reviewed here.

The early DENDRAL program was designed to derive the molecular structure of saturated acyclic mono-functional ketones [200], ethers [201], amines [47] and alcohols [203] from their low resolution mass spectra, using also their proton NMR spectra if available and if required to finally distinguish between equally possible isomers. The heart of the program is the structure generator mentioned above, and the approach can be applied to any well defined class for which general fragmentation rules are available. The program is summarised in figure 2.1, and consists of a set of class specific fragmentation rules encoded in the preliminary inference maker to determine the molecular class and the groups present (GOODLIST) and absent (BADLIST). This information can also be input if known. The structure generator then forms all possible structures, without duplication, consistent with this information. These

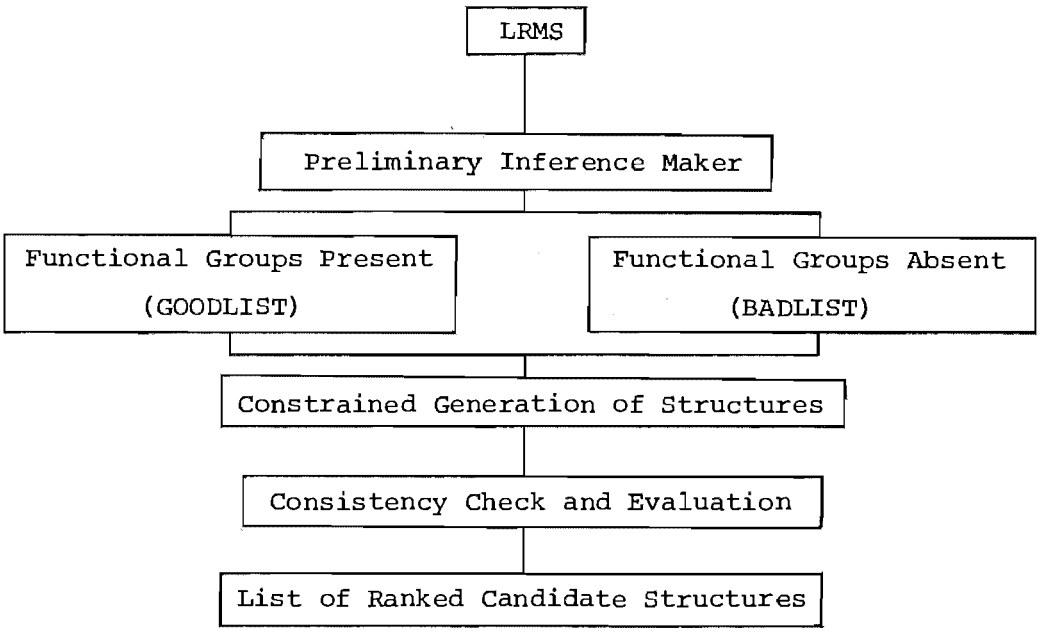


Figure 2.1: Program DENDRAL overview. Deduction of possible structures from a low resolution mass spectrum.

are used by the spectrum predictor to simulate for each its mass spectrum, using an encoded theory of fragmentation incorporating such features as likelihood of bond rupture in a given chemical environment. These predictions are then compared with the experimental spectrum, the poorest matches discarded and the others ranked in order of probability. Originally [200,201,47] the empirical formula was also required as input, but this was later [203] deduced from the highest peak and the known fragmentations of these simple mono-functionals. Program NUCL of chapter 4 corresponds to a part of the preliminary inference maker.

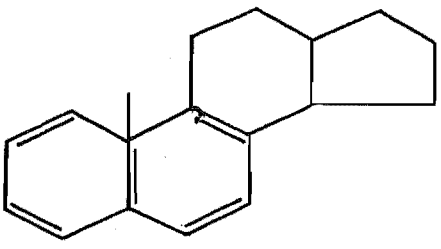
A major extension to the program was to the interpretation of the high resolution mass spectra of underivatised estrogenic steroids [205]. This demonstrated that the basic approach is not limited to very simple molecular classes so long as the fragmentations are accurately known. The identification of the unseparated components of estrogenic steroid mixtures [206] then become possible. Proof of the generality of the method was given by the incorporation in the program of an interpretation routine for the carbon-13 NMR spectra of acyclic amines [33], requiring only the empirical formula and based as for the mass spectral programs on encoded spectral rules and structure generation. Further diversification

included the development of an algorithm MOLION for general molecular weight determination [210,103] founded on the hypothesis that losses between ions in the lower half of the spectrum would reflect losses from the molecular ion. This algorithm was incorporated into the program NUCL developed as part of this present work. It is described in detail in section 4.2 where its application to nucleoside spectra is discussed.

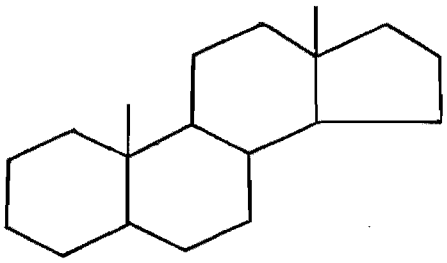
Fundamental to the Stanford project are the structure generation algorithms, both for acyclic [199] and, with more complexity, cyclic [202, 208,209,212] isomers. One application of the ability to derive a complete set of non equivalent labels for a set of cyclic isomers [209] has been in an analysis of some environmentally important fluorocarbons [211]. A further use of this exhaustive and non redundant structure generation has been in the development of the interactive program CONGEN [213]. This program is designed to assist a chemist who already has some knowledge of the structural components of an unknown molecule and wishes to investigate ways of combining them. This is the computer equivalent of the structural chemist's intuitive and often very accurate jumps from molecular information to structure [213]. One application of CONGEN was to the elucidation of the fragmentation mechanisms giving rise to, and the ionic structures generated in, the mass spectrum of triethylamine [214]. This was shown to be a surprisingly complex problem. Another application was to the identification of a tri-cyclic sesquiterpene alcohol isolated from a marine invertebrate [215], using structural components manually derived from mass spectral, IR, NMR, and carbon-13 NMR measurements.

The automatic derivation of fragmentation rules has been made possible [207,216]. When the first DENDRAL algorithms were coded, mass spectral fragmentation pathways were either culled from the literature or derived by hand, and while this is theoretically possible for all molecular classes it would soon become tiresome and repetitive. Consequently it was desired to input the spectra and structures of a set of compounds and obtain from them a list of common breakdown pathways. This was done by the program meta-DENDRAL [216] in the following steps. First, given a basic skeleton or superatom common to the set of compounds, a non-redundant list of all possible fragmentations is generated, and each structure/spectrum pair searched for evidence for each fragmentation. Hydrogen transfer is allowed for, and evidence for common fragmentation

modes is grouped together and summarised. This first part comprises program INTSUM [207], and was tested on 65 estrogenic steroids, for which the fragmentations are accurately known, and on a set of equilenins of basic skeleton I, for which they are not. The validity of the results



I



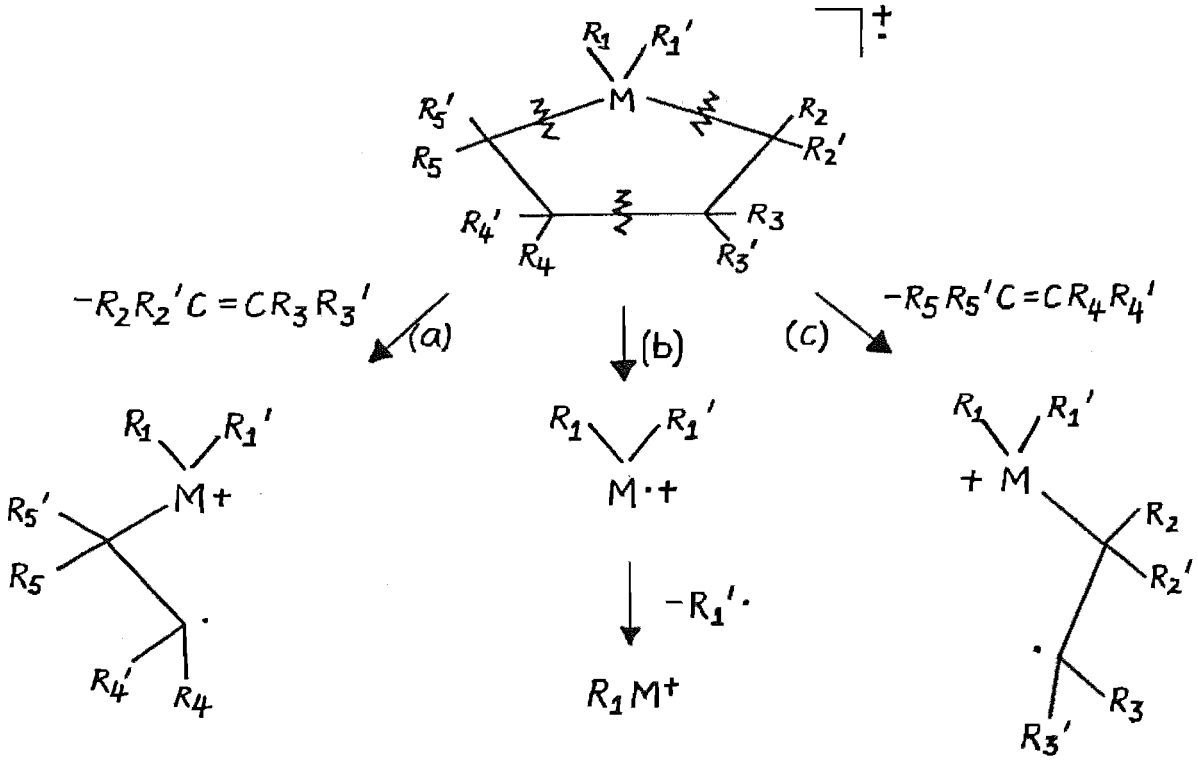
II

for the set of estrogens lends evidence to those for the equilenins. Secondly, the program meta-DENDRAL correlates the molecular processes obtained from INTSUM with the substructural environments of the bonds that are cleaved, to obtain plausible rules independent of specific environment. Thirdly, these rules are refined by merging closely related cases, restricting the application of some to specific molecular portions and generalising the application of others, and finally selecting a set of refined, general, self-consistent rules defining the mass spectra of the input compound class. The full method was again tested on estrogenic steroids, with accurate results, and applied to a previously unstudied set of mono-, di- and tri-ketoandrostanes of basic skeleton II. The validity of the rules obtained will in general depend upon the structural similarity of the input set of compounds. This restriction makes the program unsuitable for use on the present data base of nucleoside spectra (subsection 5.1.1 and appendix I) due to the structural diversity of its compounds.

2.3.7 Other Deductive Programs Several groups of workers have produced smaller programs than those outlined above, aimed either at specific classes of compounds or at a general classification of any spectrum within certain broad limits. Some of these are described in this subsection. An example of a class specific program is that of Lageot for silicon and germanium heterocyclics [217-219].

Two general classification programs are those of Smith [130,131] and Crawford and Morrison [110]. In addition, a program for the construction of fragmentation pathways devised by Delfino and Buchs [220,99] and a molecular orbital approach to mass spectral interpretation [101] are some of the many other related works which have recently been published.

The mass spectra of heterocyclics of the form and major fragmentations depicted in scheme 2.3 have been programmed by Lageot [217-219]. Given the high resolution spectrum, the molecular ion is determined assuming



M = Ge, Si

R = Me, Et, Bu, Ph, H, OH, etc

Scheme 2.3: Fragmentations of substituted silolanes and germolanes [217]

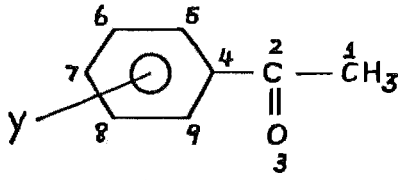
the highest peaks to be M, M-R, etc, and the fragmentation pathways (a) and (c) identified and used to reconstruct the two halves of the molecule. Pathway (b) provides further information and a DENDRAL (subsection 2.3.6) style formula notation is used to enumerate all possible structural isomers consistent with input chemical and spectroscopic

data. The program has recently been extended [219] to π-metallic complexes of such heterocyclics.

The rapid preliminary classification of components separated and analysed by GC-MS, particularly in geochemical and environmental samples, has been treated by Smith [130,131]. This program, operating on line on low resolution data, classifies each component into one of approximately fifty classes on the basis of its ion series spectrum, i.e. that set of fourteen values which are the averages of each of the fourteen homologous series. This is compared with the ion series spectra representative of each of the fifty classes, previously determined as the average of the ion series spectra of a large number of members of each class. A program with similar aims, but written for smaller monofunctional acyclics, has been described by Crawford and Morrison [110]. Low resolution spectra from GC-MS are classified as alkane, aldehyde, amine, etc on the basis of ion series spectra as above, the four largest peaks, and other features, by comparison of these with class averages determined previously from a large library data base. Class specific subroutines based [110] on known fragmentation patterns for these simple mono-functionals are then accessed. If an unknown is classified into a molecular class for which a specific sub-routine is not available, a general interrogative routine based on McLafferty's diagnostic mass peaks [221] is called.

The ion generator program of Delfino and Buchs [220,99] uses the electron book-keeping approach to mechanistic steps described by Djerassi et al. [222] to generate exhaustively and irredundantly the major primary ions present in the spectrum of any given compound. The usefulness of the program lies in that no plausible mechanism is overlooked, and the alternatives postulated can be later experimentally tested. The five mechanistic steps encoded were: ionisation with charge localisation, bond homolysis β to a radical site, bond formation between two adjacent radical sites, transfer of an atom to a radical site via cyclic transition states of various sizes, and ring closure between radical sites [99].

The numerous more theoretical approaches to the mass spectra of simple compounds are illustrated by the use of semi-empirical molecular orbital methods to calculate the effect of substituents on the intensity of the m/z 43 peak ($[COCH_3]^+$) in the spectra of a series of substituted acetophenones (scheme 2.4)[101]. The abundance of this ion was found



$$Y = \begin{cases} o - CH_3, OH \\ m - NH_2, CH_3, OH, OCH_3, CN, Br, CF_3, NO_2, C(CH_3)_3 \\ p - NH_2, N(CH_3)_2, OH, C_6H_5, OCH_3, CH_3, F, H, Cl, Br, \\ \quad CN, NO_2, C(CH_3)_3 \\ 2,4\text{-}CH_3 \end{cases}$$

Scheme 2.4: Substituted acetophenones subjected to molecular orbital calculations [101].

to depend upon the ionisation potential of the molecule and the bond density of the C2-C4 bond.

2.4 Summary of Pattern Recognition Applications

The growing number and increasingly diverse nature of problems to which pattern recognition techniques have been applied make it worthwhile grouping together those applications mentioned in this chapter. Excellent and detailed reviews of the chemical investigations reported and the methods involved have been published [223,224,140,35,185] and this summary is illustrative rather than exhaustive.

Linear discriminant analyses have been conducted for the identification of oil pollutants by neutron activation analysis [225] and IR spectroscopy [77], and for the atomic absorption spectroscopic categorisation of paper by trace metal content [226,227] as an aid to forensic science. In mass spectrometry such analyses have found uses in the prediction of spectra from coded chemical formulae [163-166], the structural elucidation of steroids [147], and the sequencing of oligodeoxyribonucleotides [1-3]. The conceptually straightforward classification method of distance from class means has been used to categorise carbon-13 NMR spectra [40] and mass spectra [143,145,146]. The nearest neighbour technique has been used to classify carbon-13 NMR spectra [40], mass spectra [150], and GC

liquid phases [19,20]. This technique was found to be the most efficacious for the classification of mass spectra in a comparison [143] with other simple methods. Linear learning machines employing error correction feedback training of decision surfaces have been described by Gray [228] and applied to phosphonate [153] and other [148,149] mass spectra. Sophisticated maximum likelihood and simplex methods have aided in the classification of binary carbon-13 NMR spectra [39e,40], mass spectra [141,142] and IR spectra [78], for which purpose the maximum likelihood approach was found superior [79]. Adaptive digital networks have also been applied to mass spectra [154, 155].

Feature extraction or dimensionality reduction has been approached by principal component analysis for the characterisation of GC liquid phases [22] and by factor analysis [156,158] for a variety of analyses. These include GC liquid phases [21], components of mixtures by GC-MS [161], and the mass spectrometric sequence analysis of oligodeoxyribonucleotides [158,162]. In addition, underlying chemical features [158] have been investigated by this technique for the relationship between structure and mass spectral mass position [160], and for the characterisation of the benzenoid isomers of $C_{10}H_{14}$ in terms of their mass spectra [157,159]. Illustrative of the burgeoning application of factor analysis, the relationship of molecular structure to metal-chelate stability for a series of diaminetetracarboxylic acids [229] has also been studied.

Chapter 3

MASS SPECTROMETRY OF NUCLEOSIDES

3.1 Introduction

Mass spectrometry is an important tool in the structural elucidation of nucleic acid components [230-233]. The spectra of the nucleosides are generally more informative than those of the corresponding free bases [231] for several reasons, and have accordingly been studied in this work. These reasons include:

(a) modifications of the sugar moiety such as 2'-O-methylation become apparent,

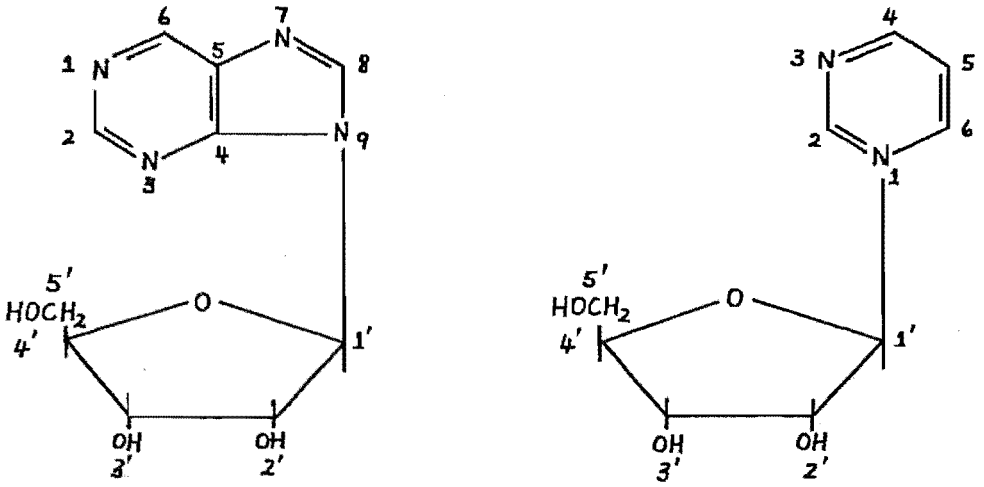
(b) the relatively more complex nucleoside spectra provide more points for tests of identity, and

(c) the spectra are somewhat more predictable than those of the corresponding purine or pyrimidine bases. This is partly because they comprise in the higher mass range a series of ions consisting of the intact base plus various portions of the sugar skeleton.

One factor that makes it more difficult to study nucleosides as opposed to the free bases is their higher polarity and correspondingly lower volatility, particularly for cytidines and guanosines. This can be overcome by appropriate derivatisation such as trimethylsilylation [234]. In this present work however, the spectra of underivatised nucleosides have been used, and the major fragmentations are described in this chapter. They form the basis of the heuristic nucleoside program NUCL of chapter 4. Modifications to these fragmentations depend upon the nature of the base and its substituents, and upon modification of the sugar. These factors for most of the common derivatives have been covered in comprehensive reviews by Hanessian [230], Hignite [233], DeJongh [232], and McCloskey [231].

Ionisation by electron impact has been most widely used for nucleosides [230-233], and although other techniques such as chemical ionisation [235], field ionisation [236,237], and field desorption [238] are receiving increasing attention [239,238d], these provide fewer points for comparison, are less widely available, and are unsuitable for the approaches of this work.

Finally, a note on nomenclature. The IUPAC numbering of purine and pyrimidine nucleosides is as follows:



The names given to the nucleosides differ slightly from those given to the free bases, and are:

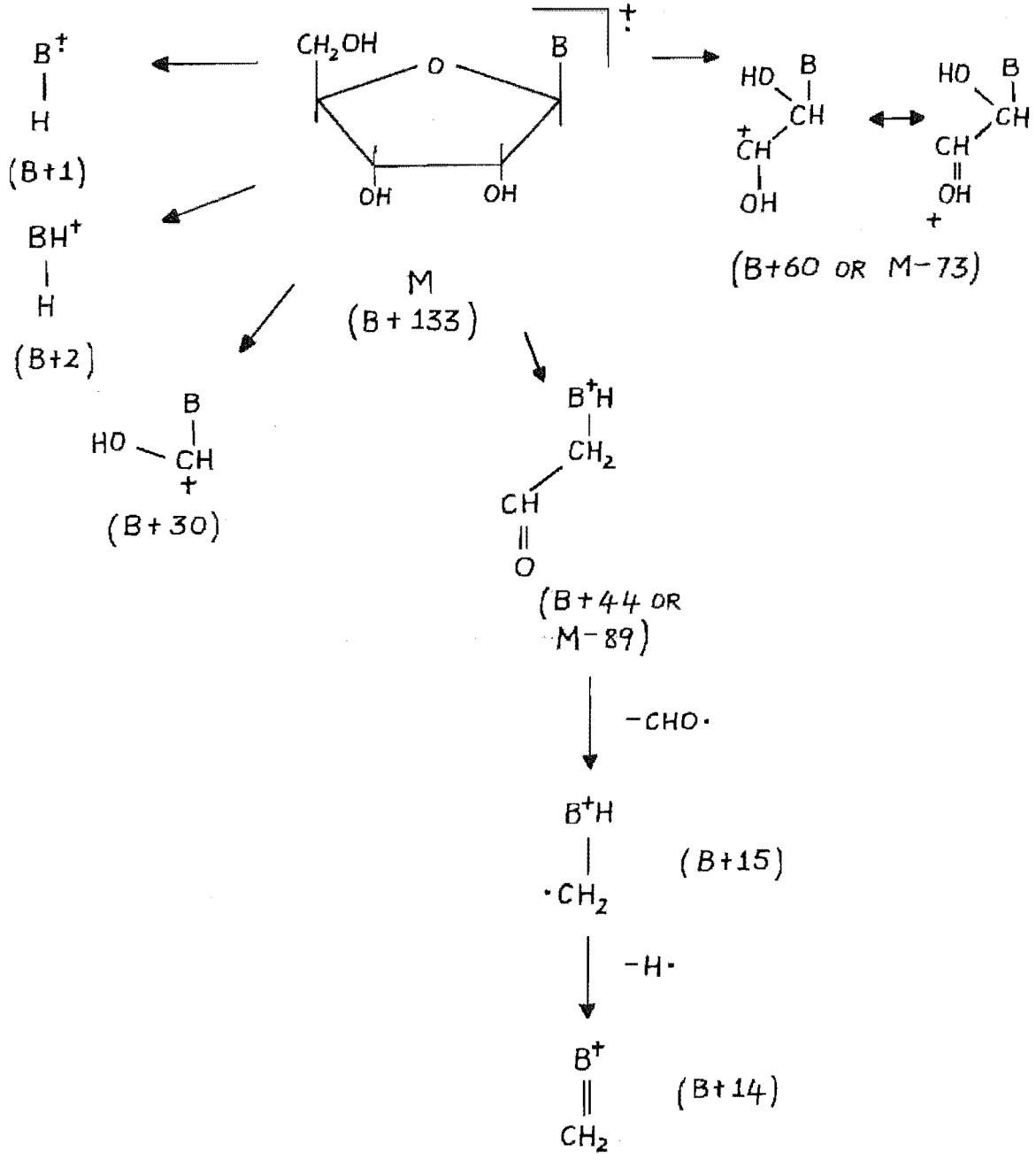
free base	nucleoside
adenine	adenosine
guanine	guanosine
cytosine	cytidine
uracil	uridine

The formula weight of the base portion is given the symbol B. In the fragmentation pathways outlined in this chapter mechanistic proofs from isotopic labelling and high resolution experiments are omitted for clarity and brevity.

3.2 Underivatised Nucleosides

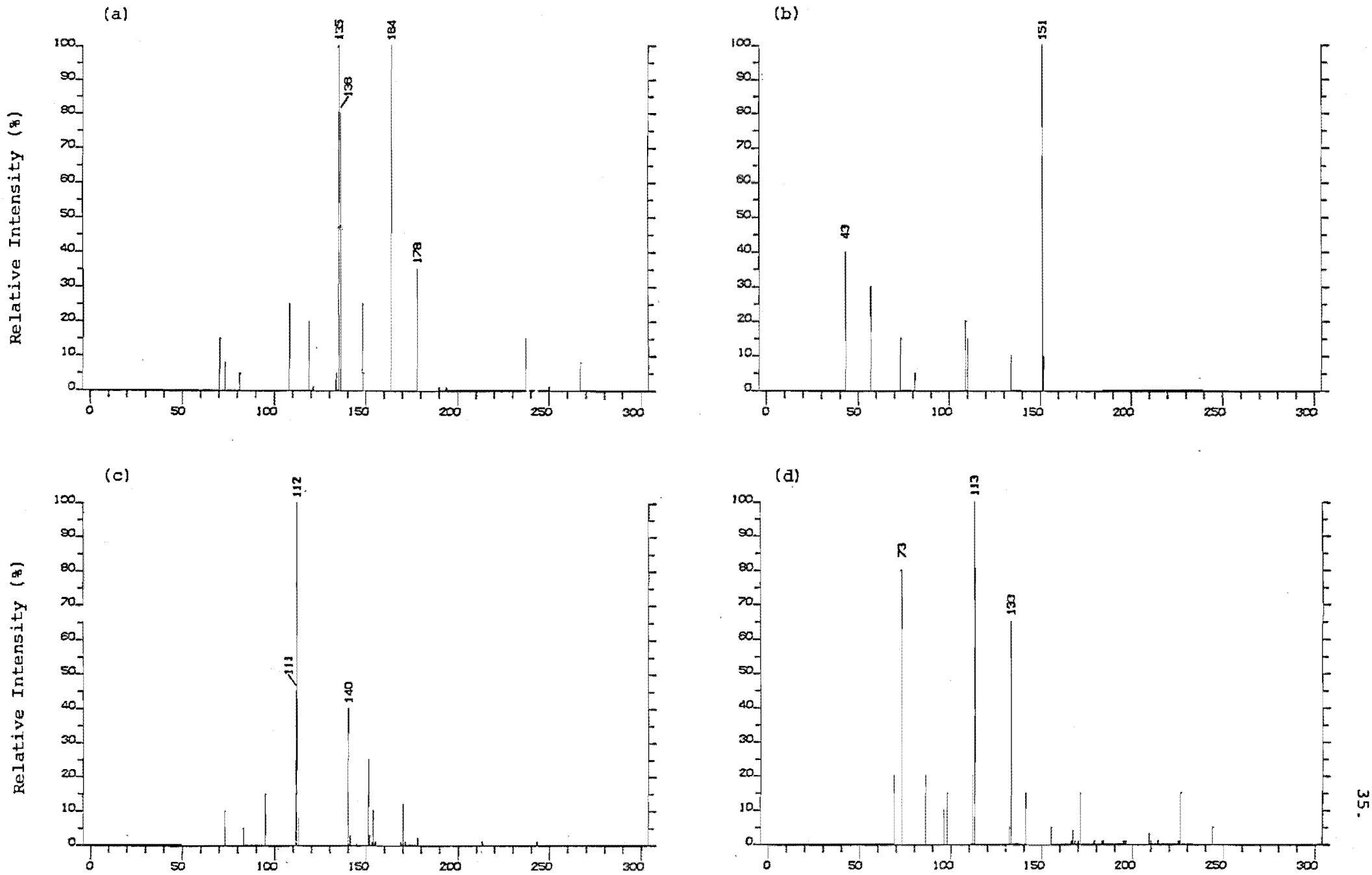
3.2.1 Common Fragmentations

Fragmentation pathways exhibited by most adenosines, guanosines, cytidines and uridines are dominated by rupture of the base-sugar bond, to form ions B+1 and B+2, and by separation of the base from the sugar with one, two or three atoms from the ribose ring attached (scheme 3.1). 70 eV electron



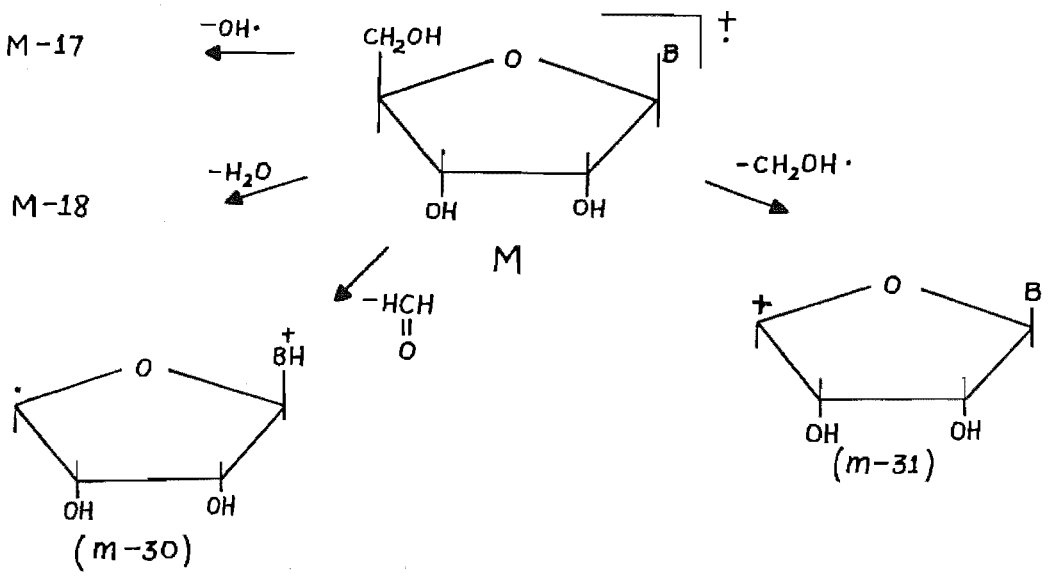
Scheme 3.1: Ions containing intact base and portion of sugar [240,231].

Figure 3.1: Mass spectra of (a) adenosine [240,243], (b) guanosine [243], (c) cytidine [243], and (d) uridine [242,243].


(a)
Relative Intensity (%)
135
184
138
178
(b)
151
43
(c)
Relative Intensity (%)
112
111
140
(d)
113
73
133
0
50
100
150
200
250
300

impact spectra of the four common nucleosides are shown in figure 3.1. A molecular ion M may or may not be present, depending in part upon the type of base (see below), and there may also be minor ions derived from M by losses of small neutral molecules or radicals from the sugar moiety. These commonly include loss of water or hydroxyl radical, elimination of formaldehyde (30 amu), or the hydroxymethylene radical (31 amu) from C-5', and various combinations of these (scheme 3.2). These ions are used in the molecular weight determination routines of program NUCL (subsection 4.2.1). Losses of water and of hydroxyl radical can occur from any of the ring hydroxyl groups [240], but many of the other ions depicted in schemes 3.1 and 3.2 have been shown [240] to be almost entirely formed by specific mechanisms. Some of the more important of these pathways are shown in scheme 3.3 for the ions B+30, B+44, B+60, and M-30, and in scheme 3.4 for the ion M-31. Structural conclusions from such fragmentations must be drawn with care, as is illustrated by the competing pathways for the elimination of the hydroxymethylene radical of scheme 3.4 for two adenosine derivatives.

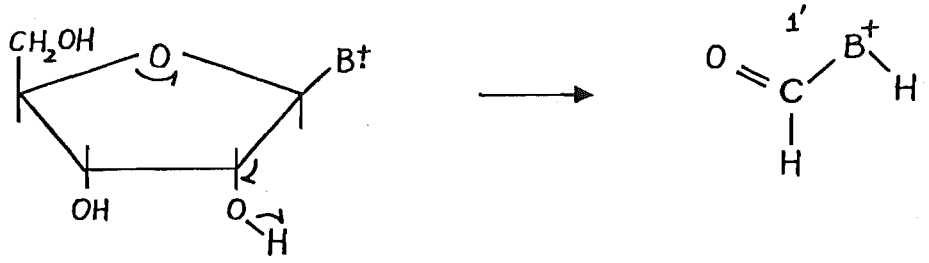
Stereochemistry can also affect the relative intensities of the various ions. For example, the M-30 ion appears with a relative intensity



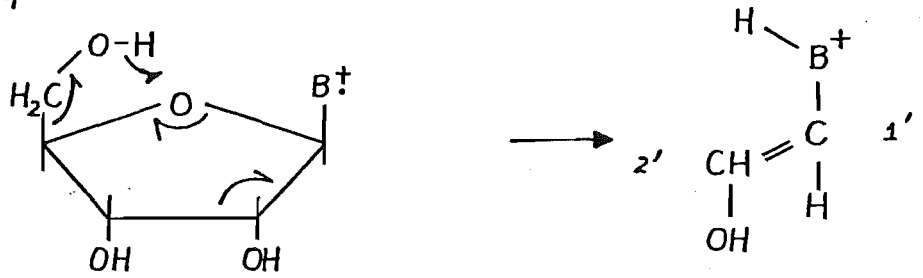
Scheme 3.2: Ions formed by small losses from M[240,231].

Scheme 3.3: Specific pathways [240]. Mechanisms for formation of the ions B+30, B+44, B+60 and M-30.

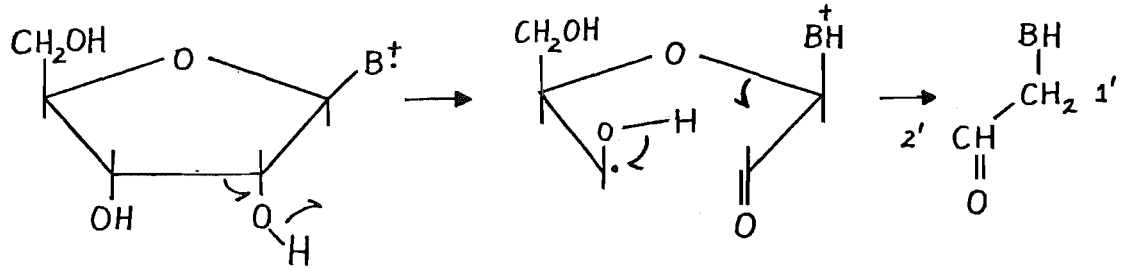
B+30



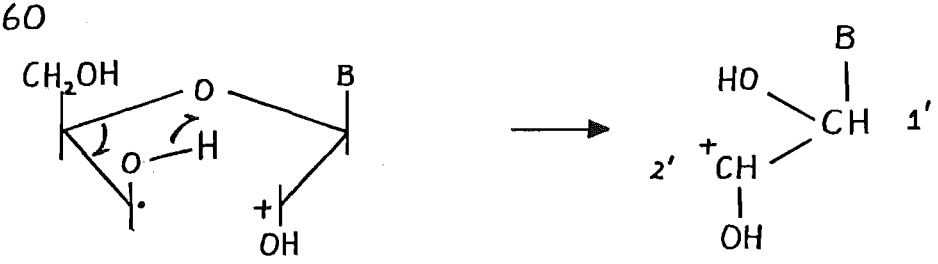
B+44



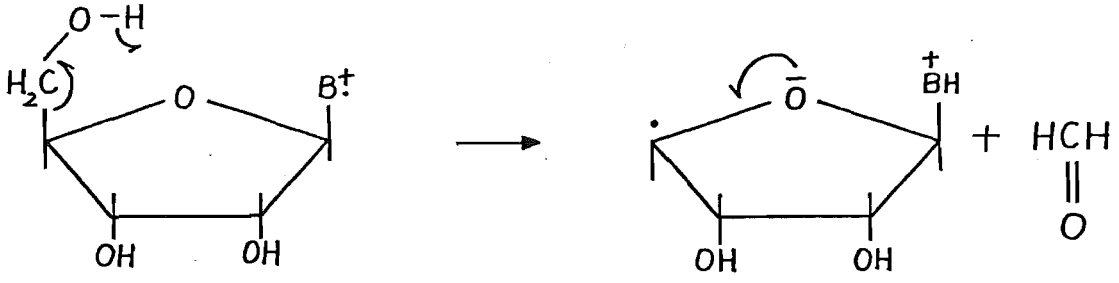
and/or



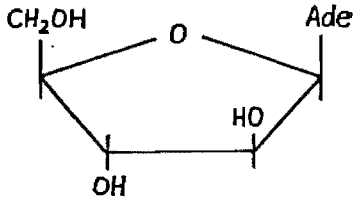
B+60



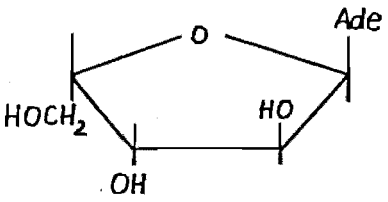
M-30



of 2.8% in the spectrum of 9-α-xylofuranosyladenine, yet is absent from that of the β anomer [240]. This follows from the mechanism for the formation of this ion postulated in scheme 3.3. In general, although



9-β-xylofuranosyladenine



9-α-xylofuranosyladenine

anomeric pairs of compounds can sometimes be distinguished by comparison of their mass spectra [240], there is no way of determining hydroxyl orientation from a single spectrum [231].

Many biologically active compounds are substituted in the sugar ring, especially at C-2', and this shifts some of the major ions to new mass values as summarised in table 3.1. As shown, ions containing C-1' retain any modification at this site, except for B+60 which disappears in 2'-deoxy compounds. This disappearance indicates that the resonance forms depicted in scheme 3.1 for the B+60 ion are

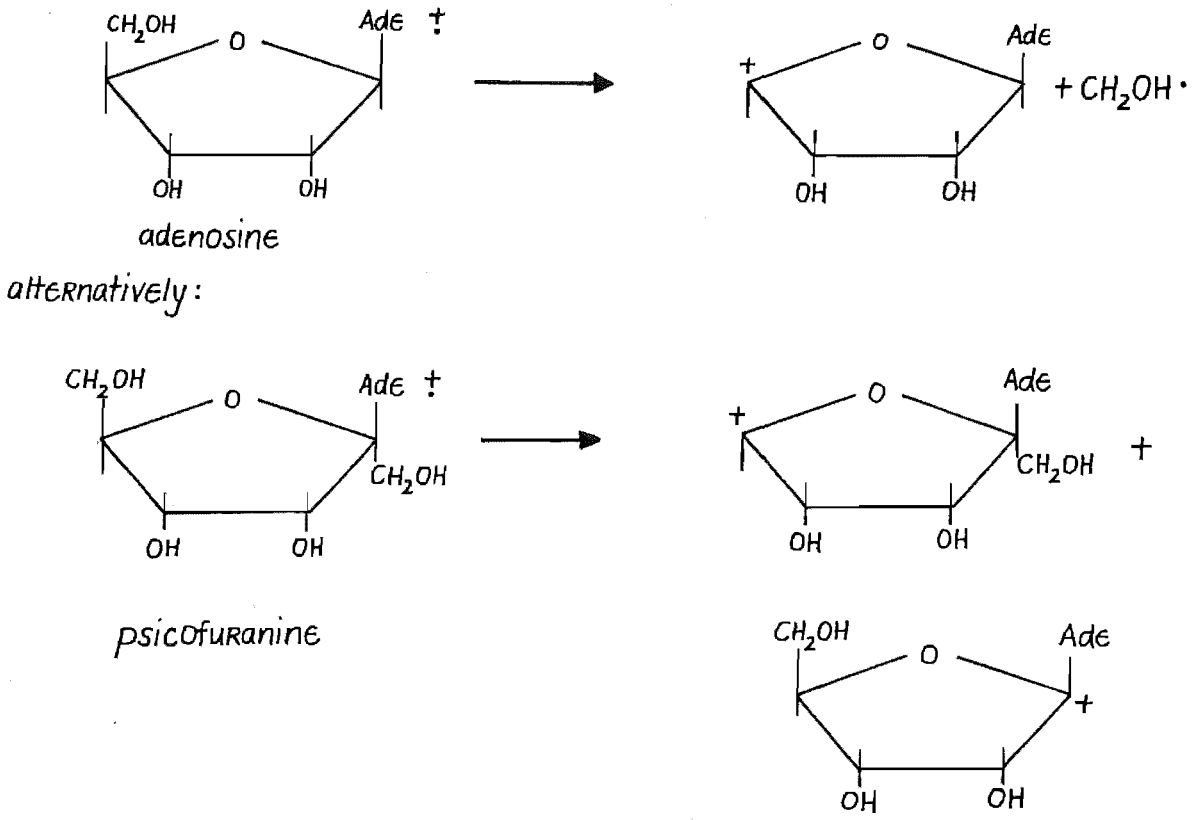
Ribose	2'-O-methylribose	2'-deoxyribose
B+30	B+30	B+30
B+44	B+58	B+28
B+60	B+74	-

Table 3.1: Effects of ribose modification. Three common base plus sugar fragment ions shifted in mass value by sugar modification.

necessary for stabilisation of the charge on the C-1' hydroxyl group. These ions, however, are not conclusive evidence for structural modification at C-2', as such compounds may also give B+44

	B+30	B+28	B+44	B+58	B+60	B+74	Reference
Adn m/z values:	164	162	178	192	194	208	
adenosine (fig.3.1(a))	100%	-	35%	-	0.2%	-	[240,243]
2'-O-methyladenosine	40	-	3	80%	1	3%	[240,241,243]
3'-O-methyladenosine	92	-	26	2	1	0.1	[240,241]
2'-deoxyadenosine	4	30%	-	-	-	-	[242]
3'-deoxyadenosine	50	-	20	-	-	-	[244]
Urd m/z values:	141	139	155	169	179	185	
uridine (fig.3.1(d))	15%	-	5%	-	15%	-	[242,243]
2'-O-methyluridine	2	-	-	2%	-	2%	[243]
3'-O-methyluridine	-	-	5	-	10	-	[243]
2'-deoxyuridine	-	2%	1	1	-	-	[242]
AdnCl m/z values:	183	181	197	211	213	217	
2'-O-methyl-6-chloro purine riboside	27%	-	3%	74%	24%	1%	[241]
3'-O-methyl-6-chloro purine riboside	95	-	26	4	7	1	[241]
2'-deoxy-6-chloro purine riboside	12	30%	-	-	-	-	[245]

Table 3.2: Relative abundances of the ions B+30, B+44, and B+60 with sugar modifications.

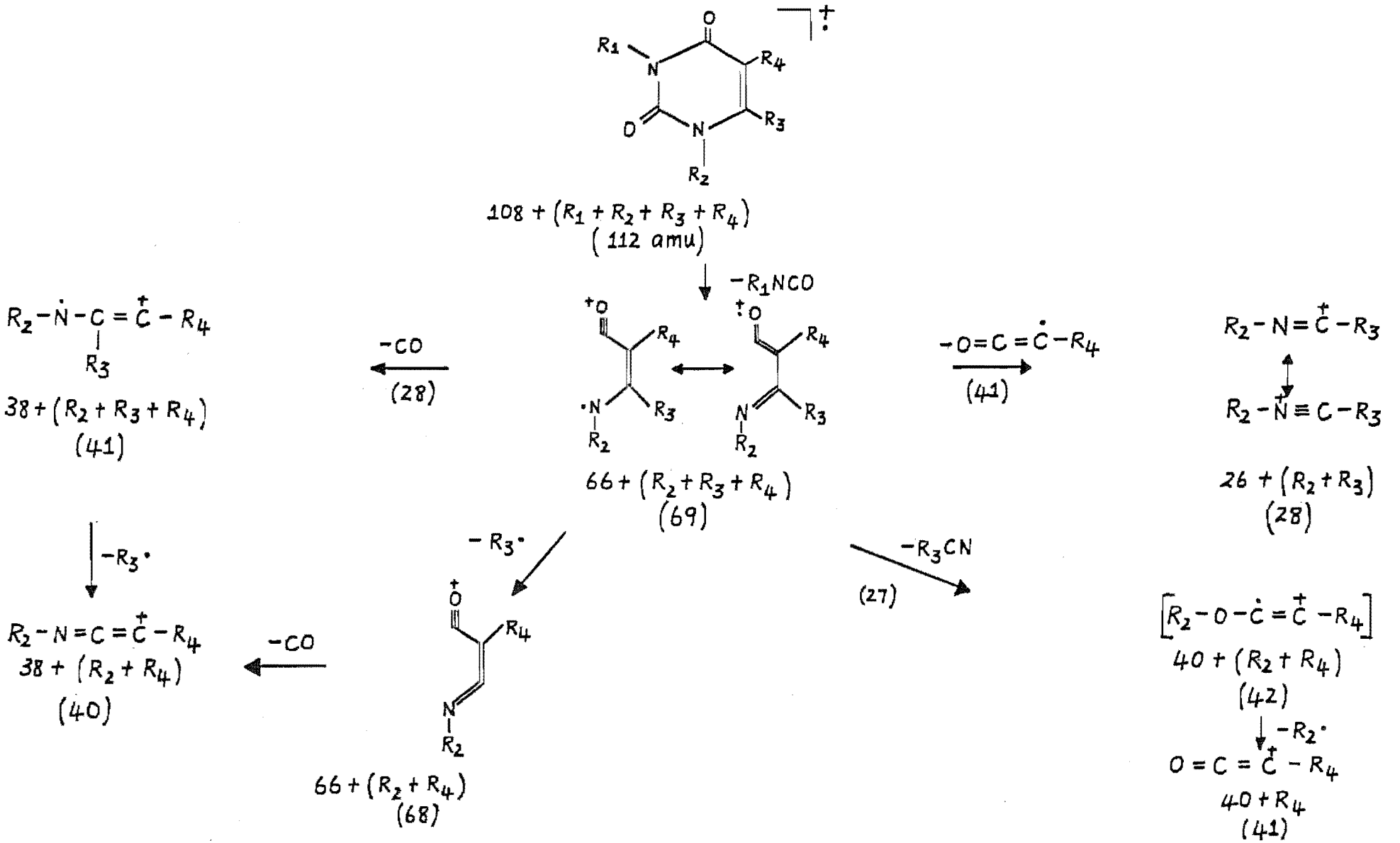


Scheme 3.4: Effect of sugar modification on M-31 ion [240]. Alternative positions from which the hydroxymethylene radical can be lost.

and B+60 ions [240,241] and nucleosides methylated at C-3' will often give B+58 and B+74 [241]. Thus several of the sugar substituted adenosines [240-243], 6-chloropurine ribosides [241,245] and uridines [242,243] summarised in table 3.2 display generally small but unexpected ions. Their weaker intensities indicate the presence of minor but competing pathways for their formation to those outlined in scheme 3.3.

A further indication of the form of the ribose ring is the presence of intact sugar ions at 133 amu for ribose compounds, 117 amu for 2'-deoxy compounds and 146 amu (sugar-1) for compounds methylated at C-2' and in some cases at C-3'. In purine nucleosides, because of the preference for cleavage with retention of the positive charge on the base, these ions are of relatively low abundance, whereas in pyrimidine nucleosides the charge is more often located on the sugar moiety and their intensity is correspondingly enhanced.

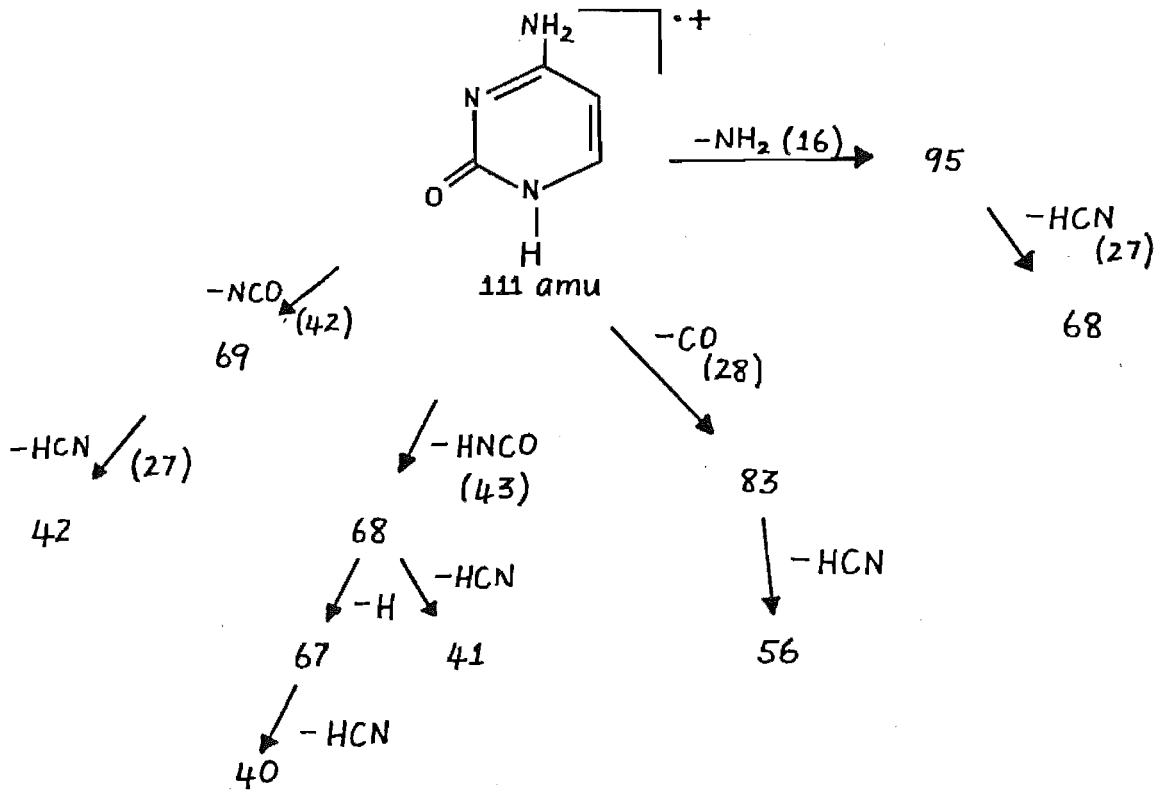
Scheme 3.5: Fragmentations of uracil derivatives [247]. Bracketed numbers refer to m/z values for uracil ($R_1 = R_2 = R_3 = R_4 = H$).


108 + (R1 + R2 + R3 + R4)
(112 amu)
-R1NCO
66 + (R2 + R3 + R4)
(69)
-CO
(28)
38 + (R2 + R3 + R4)
(41)
-O=C=Ċ-R4
(41)
R2-N=Ċ-R3
R2-Ṅ≡C-R3
26 + (R2 + R3)
(28)
-R3·
-R3·
-R3CN
(27)
R2-N=C=Ċ-R4
38 + (R2 + R4)
(40)
-CO
66 + (R2 + R4)
(68)
[R2-O-Ċ=Ċ-R4]
40 + (R2 + R4)
(42)
-R2·
O=C=Ċ-R4
40 + R4
(41)

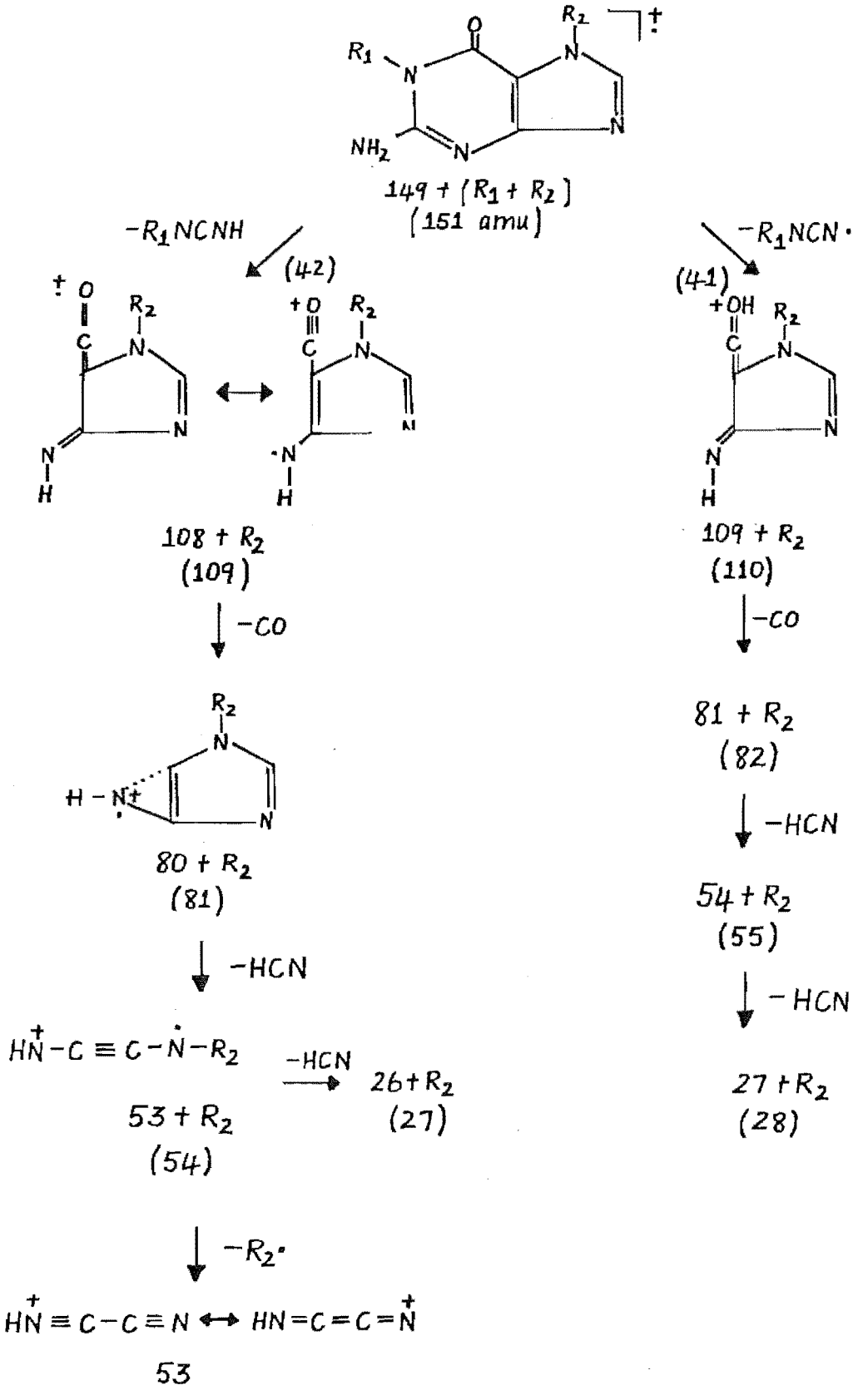
3.2.2 Base Derived Ions The ion B+1 behaves under electron impact much as the free base. Fragmentations of free pyrimidines [246-250] and purines [250-260] are well documented, and are summarised in schemes 3.5-3.8 for uracil, cytosine, guanine and adenine.

Uracil (cf. figure 3.1(d)) and its alkyl derivatives [246,247] (scheme 3.5) fragment with loss of N-3 and C-2 as R_1NCO, and thereafter lose CO, NCO, and HCN in various combinations with hydrogen or alkyl radicals. As can be seen from scheme 3.5 the resultant ions, unless highly substituted, occur at very low m/z values and so are often masked. Consequently little use has been made of them in program NUCL of chapter 4.

Cytosine (cf. figure 3.1(c)) [246,247] (scheme 3.6) loses NH_2, CO, NCO and HCN in various combinations to again give ions which occur at low mass positions. As with uracil, and also because many of the same mass values are obtained for the two bases, little use has been made of these m/z positions in the present work.



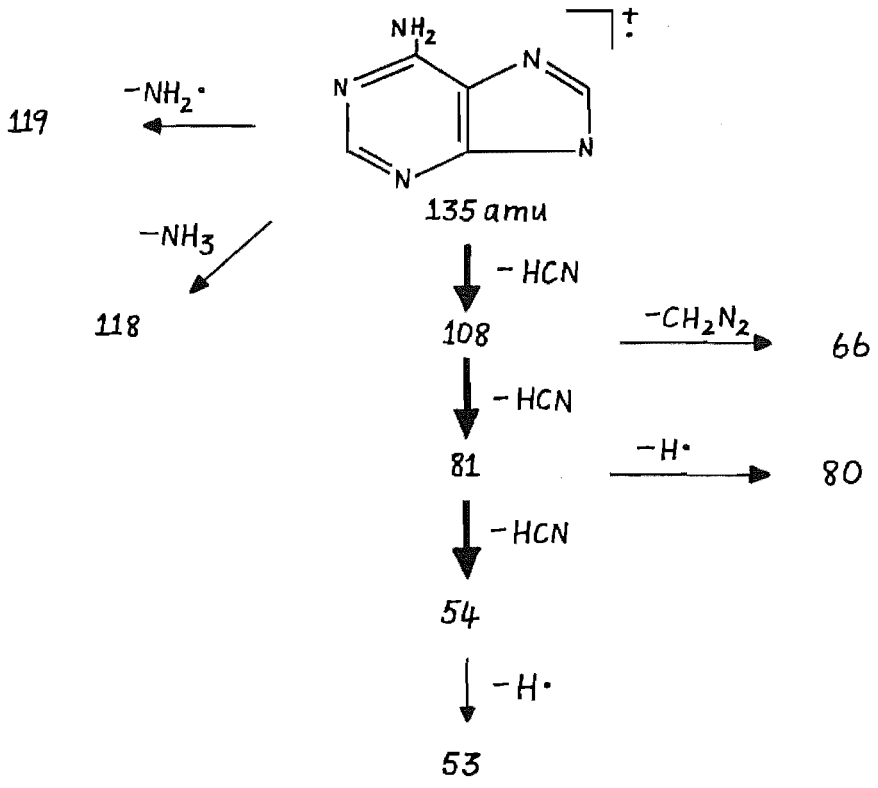
Scheme 3.6: Fragmentations of cytosine [247].



Scheme 3.7: Major fragmentations of 1- and 7- alkylguanine derivatives. 3-alkylguanines fragment analogously. Bracketed m/z values refer to guanine (R_1 = R_2 = H) [255]

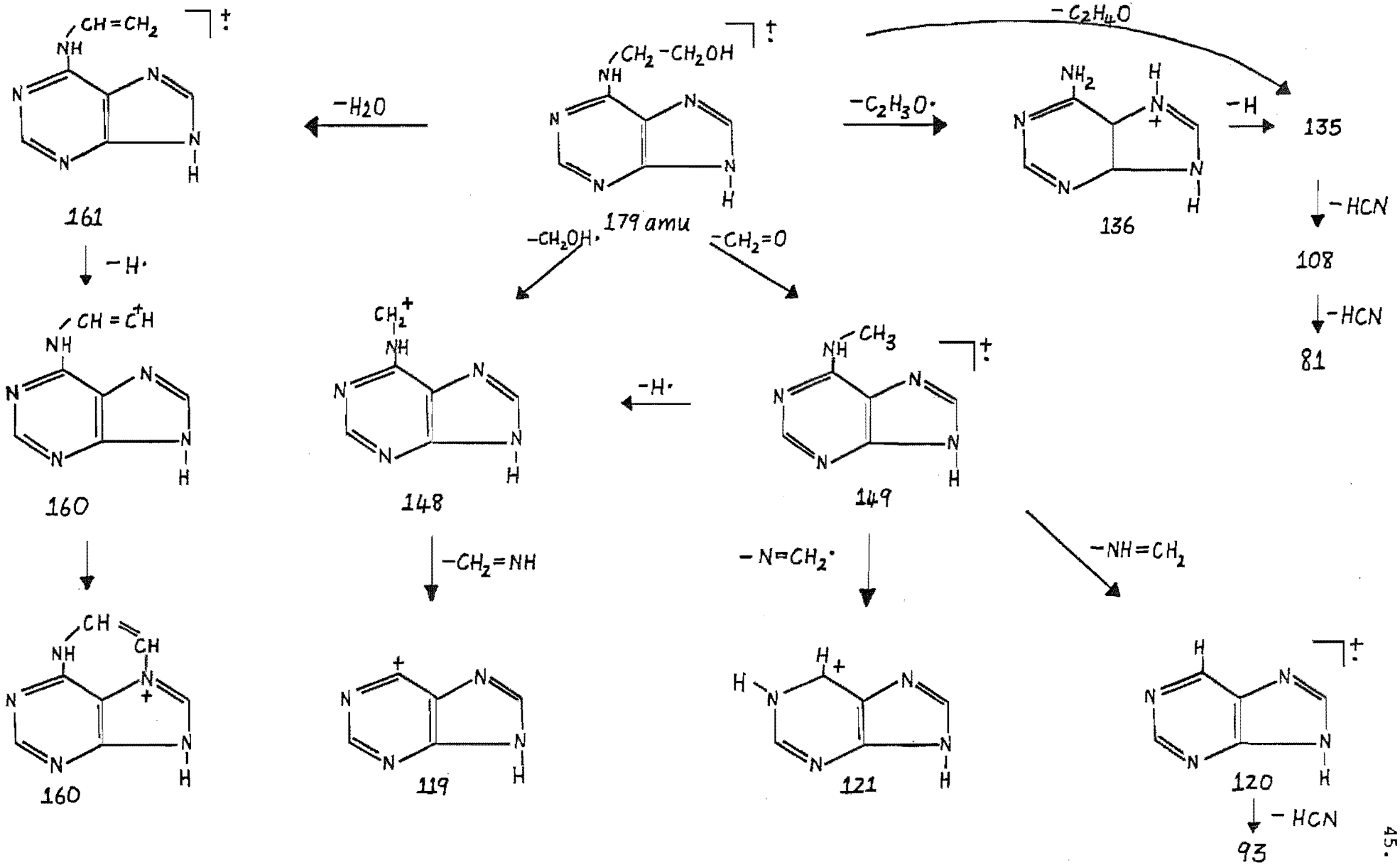
Guanine (cf. figure 3.1(b)) and its alkyl derivatives [255] are characterised by a primary loss of cyanimide, or alkyl cyanimide if substituted at N-1, from N-1 and C-2 (scheme 3.7). Secondary losses of CO and HCN originate from specific positions as indicated. Primary losses of lesser importance may also occur from M of NH_3, $CH_2N^{\bullet}$, $HNCH_2$, $R_1NH^{\bullet}$, CO, HCO, R_1NCO and $H^{\bullet}$ followed by NH_3 [255], and these are again succeeded by losses of various combinations of CO and HCN. The fragment ions considered here, although structurally diagnostic, often coincide with ions derived from uracil, cytosine or adenine (schemes 3.5, 3.6, 3.8), and consequently their usefulness in the present work is limited.

Adenine (cf. figure 3.1(a)), because of its biological significance, is the most well investigated of the four bases [250,251,253-256,258-260]. The spectrum of the parent compound is dominated by successive losses of HCN from a number of possible sites [255] (scheme 3.8), unlike the



Scheme 3.8: Fragmentations of adenine [255]. Major processes are indicated by heavy arrows.

Scheme 3.9: Fragmentations of 6-(2-hydroxyethyl)aminopurine [254].


179 amu
$-H_2O$
$-C_2H_3O\cdot$
$-C_2H_4O$
$-CH_2OH\cdot$
$-CH_2=O$
161
160
160
148
149
136
135
108
81
119
121
120
93
$-H\cdot$
$-CH_2=NH$
$-N=CH_2\cdot$
$-NH=CH_2$
$-HCN$
$-HCN$
$-HCN$

highly specific fragmentations of guanine, and structural information from these is minimal. This absence of specificity notwithstanding, the ions at 119,108 and 81 amu are characteristic of adenine and occur also in the spectra of biologically active N-6 substituted derivatives. Such substitution can significantly alter the fragmentation scheme [254,255] as now the N-6 side chain bond is often of a comparable lability to the ribose-base nucleosidic linkage. Consequently, in the spectra of N-6 substituted adenosines the two primary decay paths are often loss of the side chain from both the base B+1, and the molecular ion, B+ sugar. These follow the pattern outlined in scheme 3.9 for 6-(2-hydroxyethyl)aminopurine[254], and yield ions characteristic of N-6 substitution at 149,148,135,121,120,119,108 and 81 amu.

3.2.3 Carbon Glycosides Nucleosides with a carbon-carbon glycosidic linkage exhibit distinctive and characteristic mass spectra [231, 261-265, 241]. The enhanced stability of the linkage renders the intensity of the B+1 and B+2 ions small or negligible, and the B+30 ion is generally [264] the most abundant. These features are illustrated by the representative spectra of pseudouridine (figure 3.2(a) and scheme 3.10) and formycin (pseudoadenosine) (figure 3.2(b)). The ions depicted in scheme 3.10 have been utilised in the determination of carbon glycoside mass by program NUCL (subsection 4.3.1).

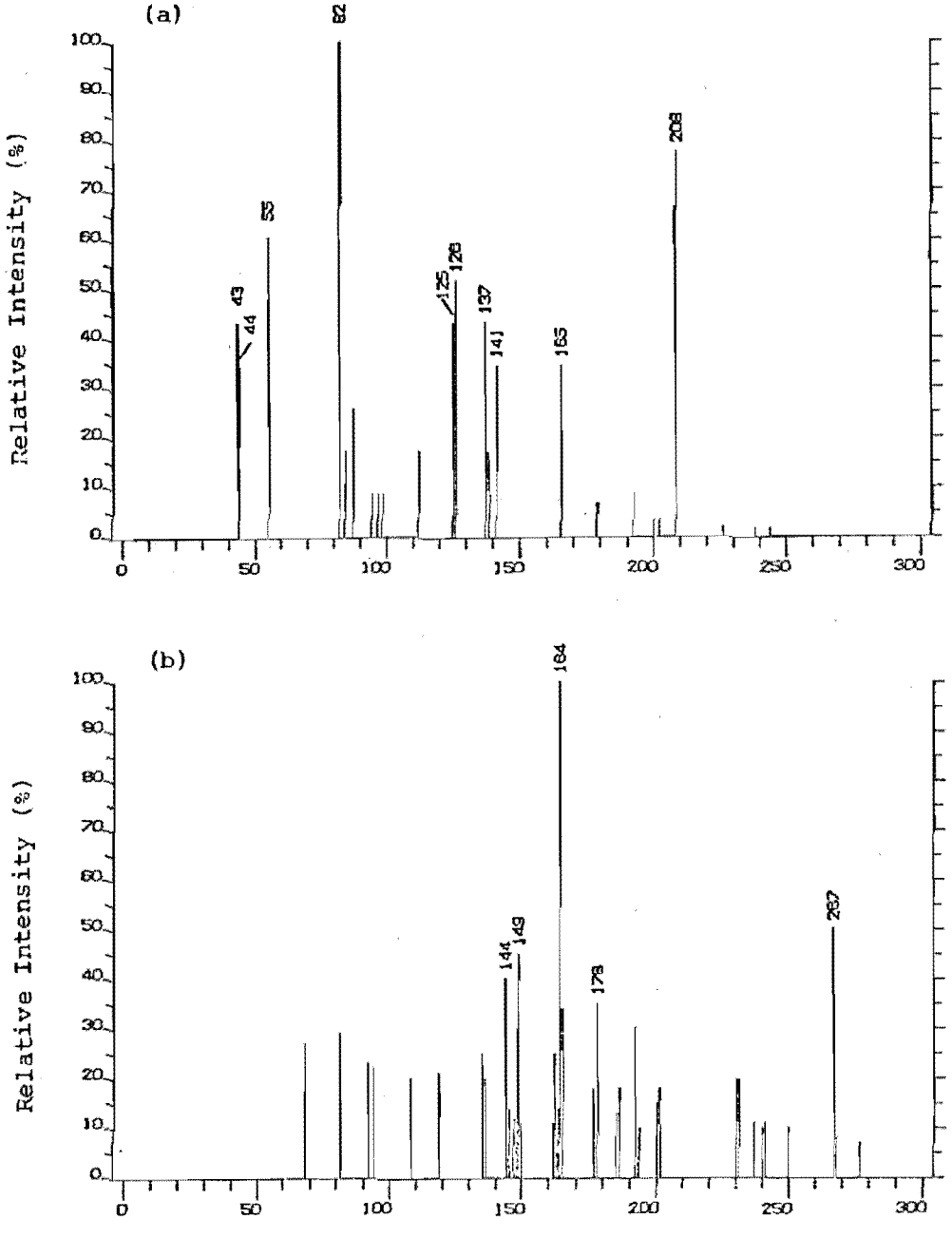
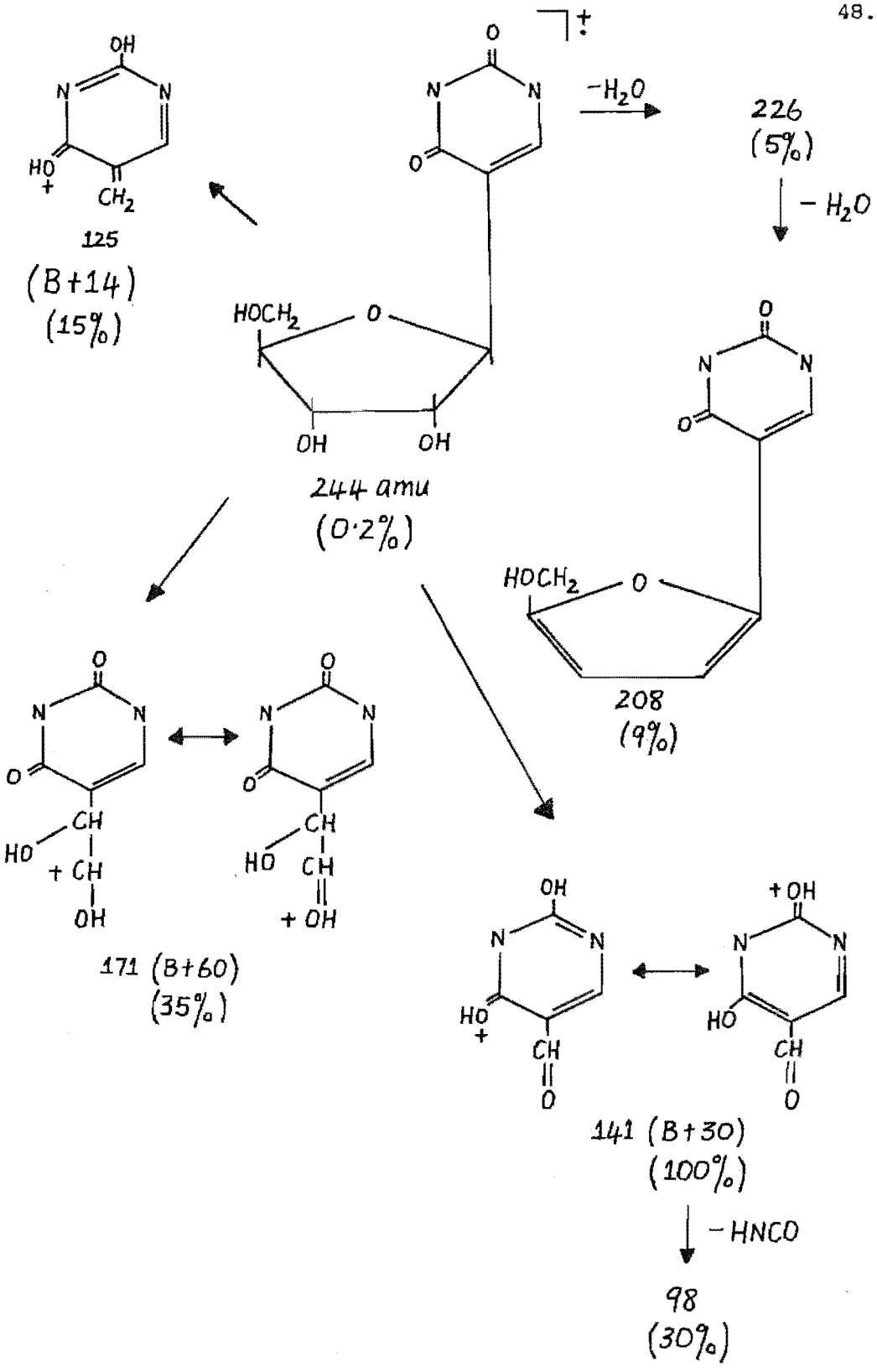


Figure 3.2: Mass spectra of (a) pseudouridine [261] and (b) formycin (pseudoadenosine) [263].



Scheme 3.10: Fragmentations of pseudouridine [231,261,262]. Relative intensities given in brackets.

CHAPTER 4

HEURISTIC FRAGMENTATION PROGRAM

4.1 Introduction

A heuristic mass spectral interpretation program has encoded into it the characteristic fragmentations of a certain class of compounds. Those for underivatised nucleosides described in chapter 3 were encoded into program NUCL in order to elicit from an unknown nucleoside mass spectrum major features of its structure. Because of the wide diversity possible amongst nucleosides, as evidenced by the composition of the data base (appendix I), complete identification by this approach is not at present possible. This fact, together with the short process time required for a single compound (section 4.5), makes this program perhaps most suited to GC-MS analysis of a mixture of nucleosides. In such a system only rapid preliminary identification would be required and more complete structural determination could later be performed manually and in conjunction with the pattern recognition approaches of chapters 6-9. To assist manual interpretation all significant ions and ion series found in an unknown spectrum were included in the output listing (see appendix II).

Identification of three major features was attempted: the molecular weight M, the base weight B, and the general nature of the base. In addition, a more general approach to molecular weight determination embodied in the externally supplied program MOLION (section 4.2) is compared with the heuristic method. The first two features of structure mentioned above could be determined quite satisfactorily, although before use as an on line identification tool some improvement could perhaps be effected. The third feature, the nature of the base, could not however be successfully deduced from the spectrum by this method.

4.2 Molecular Weight Determination

Three approaches to the determination of the molecular weights of underivatised nucleosides were investigated. The externally supplied program MOLION gave marginally better overall performance than two specially written subroutines which, however, surpassed it for certain compound types.

4.2.1 Methods The two algorithms written as part of program NUCL are conceptually slightly different from program MOLION. The latter uses the lower half of the spectrum to deduce primary losses from the molecular ion M. The former look for known nucleoside primary losses, from M candidates postulated in various ways. The three methods are as follows.

(1) Program MOLION, a part of the heuristic DENDRAL project of Stanford University, was written and supplied by R.G. Dromey [103b] and is fully documented elsewhere [210]. The method is based upon the postulate that

> "there exists at least one 'secondary loss' in a spectrum that will match a 'primary loss' from the molecular ion irrespective of whether the molecular ion peak is present in the spectrum"[210].

A primary loss for the purposes of the program is taken as one from a postulated molecular ion, and a secondary loss as one from any other ion together with all masses in the lower half of the spectrum. This assumption is tantamount to saying that

> "any mass lost from the molecular ion will also be observed as a loss from one fragment ion to another, or will itself appear as a fragment ion in the lower regions of the spectrum"[210].

Certain chemical restrictions, such as a set of unlikely primary losses, are also imposed.

(2) Losses from M. Two procedures written as parts of program NUCL specifically for underivatised nucleosides were based on a set of common losses from the molecular ion [266]. Each peak in turn from the highest peak (HP) down to HP-61 was considered to arise from a possible molecular ion by way of one of the following losses from the sugar portion:

M		
M-15	(CH_3)	
M-17	(OH)	
M-18	(H_2O)	
M-30	(CH_2=O)	
M-31	(CH_2OH)	(4.2.1)
M-32	(CH_3OH)	
M-35	(H_2O + OH)	
M-36	(H_2O + H_2O)	
M-61	(CH_2=O + CH_2OH)	

Losses of 30,31 and 32 generally occur by cleavage of the C-4' - C-5' bond in ribose sugars, unless one or more of the hydroxy groups is methylated in which case these losses occur preferentially from such sites [240]. Mechanistic details are supplied in subsection 3.2.1. Losses are considered only from the sugar moiety because these are reasonably constant over a wide variety of nucleosides. Conversely the structure of the base varies so greatly that many different primary losses may occur. To allow for small peaks at M+1 etc, molecular weight candidates are considered if they are of greater mass than the third highest peak in the spectrum. Each potential candidate is checked for the same ten losses as above, and if another is present (i.e. making two such ions in total including the peak from which the candidate was postulated) that value is included as a molecular weight candidate. These are ranked according to the sum of the mass (m_i) x intensity (I_i) values of the ten ions above. This product

$$0.001 \sum_i m_i I_i \qquad (4.2.2)$$

summed over the various evidential ions is used throughout program NUCL to rank the various candidates in order of likelihood.

(3) Losses from B + Sugar(S). The third method differs from the second only in the means of formation of the potential molecular weight candidates. The masses of each of the postulated base candidates, obtained as described in section 4.3, are added to each of the three most common ribose sugar masses 117, 133, and 147 amu, viz. deoxyribose, ribose, and methylribose. This approach has the advantage of being directly related to and providing a check upon the independently derived base masses (section 4.3), but will obviously fail for any more highly modified sugar moieties than the three most common. Only one of the ten primary losses listed in equation (4.2.1) above is required to confirm a particular mass value as a candidate.

4.2.2 Presentation of Results The percentage successes obtained on the set of 125 nucleoside spectra (appendix I) are presented for each of the three approaches in table 4.1 and are graphed in figures 4.1 and 4.2. Two criteria for judgement are presented: ranking of the correct molecular weight (a) as the first,

Table 4.1: Molecular weight determination. Correct values ranked first ("1") and amongst the top five candidates ("5") for seven structural categories, subdivided according to sugar type. These are D-ribose ("133"), 2'- or 3'- O-methyl or 2'- or 3'- deoxy ribose ("2/3 m/d"), these common types combined ("3 comm"), all other sugar forms ("Oth"), and all sugars ("Tot"). Three methods used; program MOLION, program NUCL using losses from postulated molecular weights ("M"), and program NUCL using losses from postulated base weights plus formula weights of three common sugars ("B+S"). Numbers of compounds in each structural category shown in brackets.

		MOLION			M			B+S	
		1	5		1	5		1	5
Adn(all)	133	65%	78%		48%	70%		74%	87%
	2/3 m/d	60	100		80	96		76	96
(74)	3 comm.	63	90		65	83		75	92
	Oth	73	81		35	54		4	8
	Tot	66	87		54	73		50	62
Adn (Sug.	133	-			-			-	
modif.)	2/3 m/d	60%	100%		80%	96%		76%	96%
	3 comm.	60	100		80	96		76	96
(51)	Oth	73	81		35	54		4	8
	Tot	67	90		57	74		39	51
Adn (N6	133	67%	78%		39%	61%		67%	83%
modif.)	2/3 m/d	61	100		72	94		67	95
	3 comm.	64	89		55	78		67	89
(50)	Oth	79	79		36	50		0	0
	Tot	68	86		50	70		48	64
Pur	133	64%	77%		50%	68%		73%	86%
	2/3 m/d	52	100		80	96		76	96
(74)	3 comm.	57	89		66	83		73	92
	Oth	70	78		33	56		7	11
	Tot	62	85		54	73		50	62
Pyr	133	28%	100%		78%	78%		89%	89%
	2/3 m/d	89	100		89	89		55	78
(26)	3 comm.	56	100		83	83		72	83
	Oth	87	100		50	75		37	37
	Tot	65	100		73	81		61	69
C-Glyc	133	13%	50%		50%	100%		37%	87%
	2/3 m/d	50	87		50	87		37	27
(20)	3 comm.	31	69		50	94		38	62
	Oth	50	75		0	0		0	25
	Tot	35	70		40	75		30	55
All	133	45%	78%		55%	78%		70%	87%
	2/3 m/d	60	98		78	93		67	82
(125)	3 comm.	53	88		67	86		68	85
	Oth	70	80		32	55		12	17
	Tot	58	86		56	76		50	63

most likely, candidate in figure 4.1, and (b) amongst the top five most likely candidates in figure 4.2. A detailed tabulation of the individual behaviour of each of the 125 spectra with each of the three methods, from which this table and these graphs have been constructed, is presented in appendix III.

The spectra have been classified according to both sugar type and nucleoside structure. In graphs (a) of figures 4.1 and 4.2 those 40 nucleosides with D-ribose (formula weight 133 amu) as their sugar moiety are presented. In graphs (b) are those 85 with the three most common sugar forms: D-ribose (133 amu), 2'- or 3'- O- methylribose (147 amu), and 2'- or 3'- deoxyribose (117 amu). In graphs (c) are those 40 with all other sugar forms, these are generally but not always substituted riboses, and in graphs (d) are the total i.e. the sum of the values represented in (b) and (c). These four sugar types apply also to figures 4.3 and 4.4.

Seven nucleoside categories are plotted on an arbitrary scale along the ordinate in each graph:

(a) total adenosines (Adn)
(b) adenosines with a modified, i.e. non D-ribose, sugar (Asug)
(c) adenosines modified at the N-6 position (AN6)
(d) total purines (Pur)
(e) total pyrimidines (cytidines and uridines) (Pyr)
(f) carbon glycosides (CG)
(g) total nucleosides (Tot).

The first four exhibit a high degree of overlap. The last, all inclusive category is the most significant.

It is worth mention that of the 125 spectra, all except five exhibit molecular ions and often M+1 etc peaks as well. The five are a uridine, spectrum 26 (as numbered in appendix I) of highest peak M-18, a cytidine, spectrum 40 (M-15), a carbon-glycoside, spectrum 48(M-59), an 8-oxyadenosine, spectrum 54 (M-16), and an N-6 substituted adenosine, spectrum 100, which loses part of its N-6 side chain to give a highest peak at M-75.

4.2.3 Results and Discussion The correct molecular weight was ranked first in only 50-58% of the cases, varying slightly with method, as is evident from the last three data points of figure 4.1(d) for all varieties of structure. The correct value was however

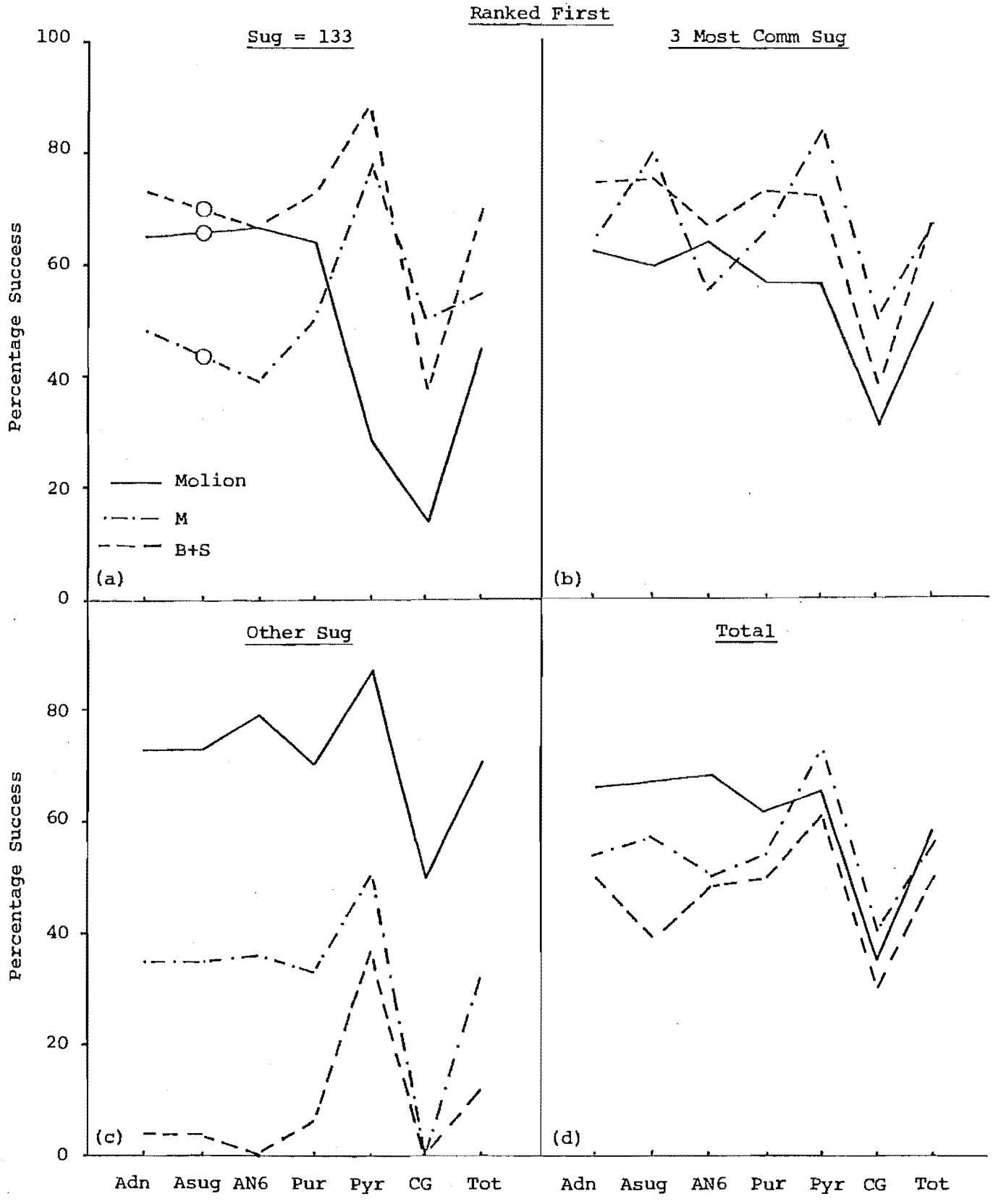


Figure 4.1: Molecular weight determination: correct candidate ranked first. Percentages of seven overlapping structural categories (subsection 4.2.2) for which the correct molecular weight was ranked the most likely candidate, using three different algorithms. Structural categories further divided according to sugar type (a)-(d) (subsection 4.2.2).

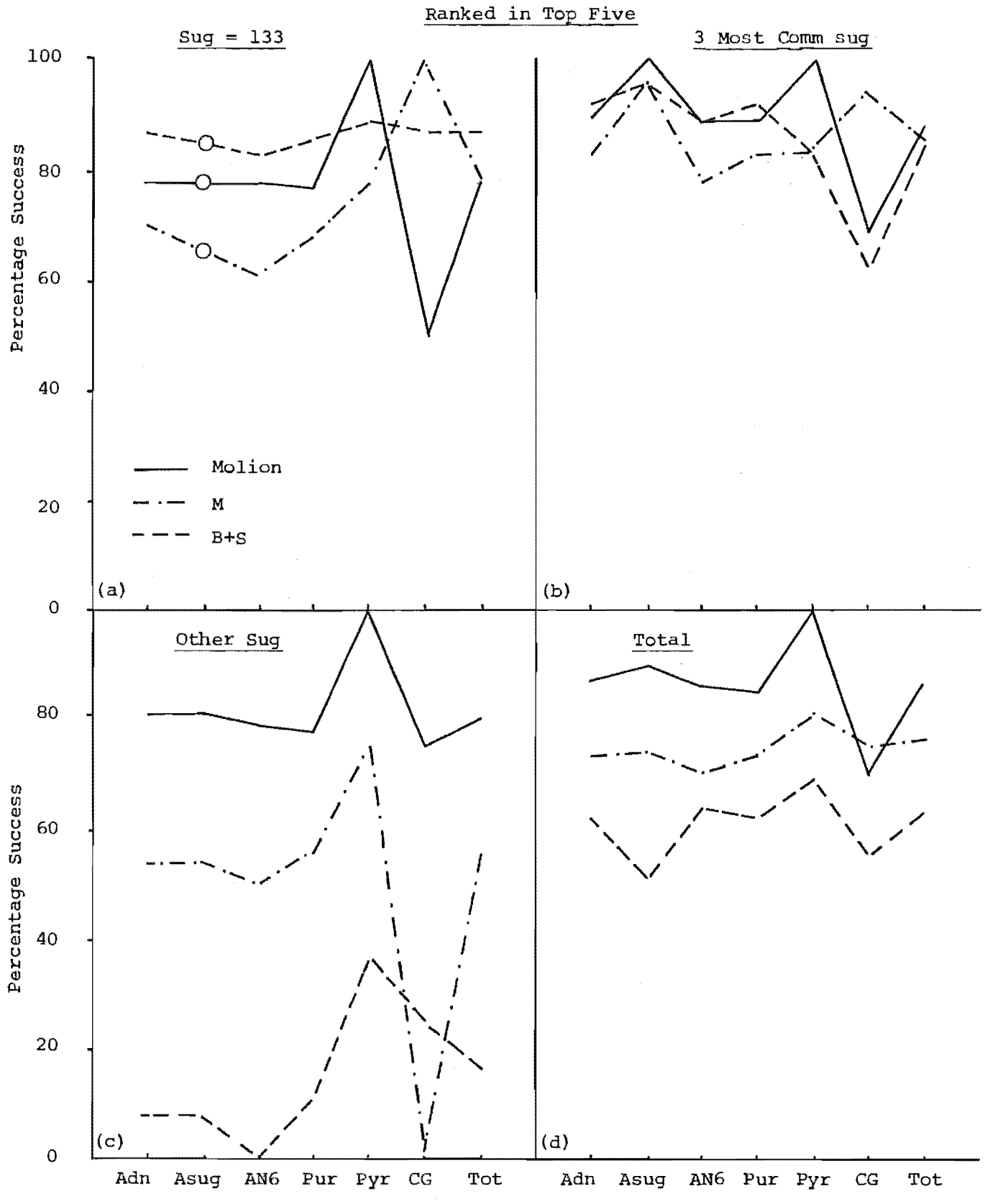


Figure 4.2: Molecular weight determination: correct candidate ranked amongst the first five. Percentages of seven overlapping structural categories (subsection 4.2.2) for which the correct molecular weight was ranked amongst the first five most likely candidates, using three different algorithms. Structural categories further divided according to sugar type (a)-(d) (subsection 4.2.2).

ranked amongst the top five candidates in 63%-86% of the cases (figure 4.2(d)). This latter figure of 86% for MOLION makes it clearly the best performed method if the criterion adapted for judgement is that of inclusion in the top five candidates (compare the three graphs in figure 4.2(d)). If however it is required that the correct value actually be ranked first, then there is only marginal distinction between the three methods. This can be seen from the three graphs of figure 4.1(d). MOLION performs the best of the three methods on adenosines and purines in general (the first four points of the graphs) but when all spectra are considered (the last point) the three methods differ by under 10%. Selected features of these results will now be examined more closely, dealing both with effects of structure and with relative performance of the three methods.

The four highly overlapping purine categories forming the first four data points of each graph: total purines, total adenosines, and the two types of substituted adenosine, generally behaved, as expected, very much alike. Relative to these categories, pyrimidines were almost invariably easier to identify, and carbon glycosides almost invariably more difficult. With these exceptions there are few consistent trends amongst structural categories in the twenty-four graphs of figures 4.1 and 4.2. It is however worthwhile noting that carbon glycosides become very tractable to the losses from M routine of program NUCL if the sugar moiety is one of the standard types (figure 4.2(a)-(b)). In such cases the greater strength of the C-C glycosidic bond compared to the C-N nucleosidic linkage leads to increased stability of the higher mass range ions and hence often renders the hydroxy groups of the sugar the most labile sites. Consequently the fragmentations at these sites required by the routine occur with perhaps greater reliability than with other forms of nucleoside.

All these methods performed approximately equally well on the most common sugar forms (figures 4.1(a)-(b) and 4.2(a)-(b)). As expected, given the specificity of their algorithms, the two routines of program NUCL operated less well on more varied structures (figures 4.1(c) and 4.2(c)). The losses from B+S routine in particular, the more specific of the two, fared poorly on non standard sugars (figure 4.2(c)). On the other hand, it performed at least as well as the other methods when applied to the sugar types for which it was designed (figures 4.1(b) and 4.2(b)) and was superior to them on D-ribose type sugars (figures 4.1(a) and 4.2(a)) alone. The losses from M routine also performed poorly on non standard sugars (figure 4.2(c)). This could only be expected given that

many of the common losses involve the hydroxyls of the ribose and that these were often inaccessible in such sugars.

Finally, a note on the performance of the three methods on those spectra lacking a molecular ion. The correct molecular weight was placed by MOLION in the top five candidates only for spectra 26 and 40. Both parts of program NUCL succeeded only on spectrum 26, but at least ranked the correct value first. They failed on spectra 40, 48, 54 and 100 whereas MOLION failed for spectra 48, 54 and 100. Of these nucleosides (see appendix I) three (26, 40 and 100) contained a D-ribose sugar moiety, while 48 and 54 were both acetylated in the sugar. The failure of NUCL in both these latter cases is understandable, as the program was constructed solely for underivatised nucleosides.

4.3 Base Weight Determination

Procedures relying upon fragment ions composed of the intact base plus portions of the sugar (section 3.2) were written to determine the formula weight (B) of the base part of underivatised nucleosides and carbon glycosides. The methods, although not as yet of high enough reliability for on line use, appear promising. The correct mass is generally (~ 85%) ranked within the top five candidates, although it is seldom (~ 57%) placed first.

4.3.1 Methods These two procedures of program NUCL utilise intact base plus sugar portion ions, and hence are essentially independent of the nature of the base except in so far as this affects the abundances of such ions.

(1) Nucleoside base candidates. This approach is based upon the general nucleoside fragmentations described in subsection 3.2.1. Each ion in the spectrum of 10% or greater relative abundance is considered as a possible B+1 or B+2 ion. If B+1 is of 40% or greater relative intensity, or if at least three of the six ions listed below are present with one of at least 20% relative intensity, that value of B is selected as a base candidate. The relevant ions determining selection are:

B
B+1
B+2
B+30 (4.3.1)
B+44 or B+58 or B+28
B+60 or B+74

The last two of these may vary depending on whether the sugar ring is methylated at the C-2' position (B+44 and B+60 become B+58 and B+74, respectively), or is lacking a hydroxyl group at this position (B+44 becomes B+28 and B+60 disappears) as described in subsection 3.2.1. The candidates are ranked according to the value of

$$0.001 \sum_i m_i I_i \qquad (4.3.2)$$

for the ions of mass m_i: B, B+1, B+2, B+14, B+15, B+28, B+30, B+44, B+58, B+60, and B+74.

(2) Carbon glycoside base candidates. As described in subsection 3.2.3, the ion B+30 is almost invariably [264] dominant in the spectra of carbon glycosides, and this part of program NUCL relies heavily upon this ion. Each peak above m/z 109 of 20% or greater relative intensity is considered as a possible B+30 candidate, and confirmation is provided if the ions B+1 and B+2 are both not more than 30% relative intensity and at least one of the following ions is present:

B+14
B+15
B+28
B+44 (4.3.3)
B+58
B+60
B+74

The candidates are again ranked according to the sum given by equation (4.3.2) for these seven ions plus B+30.

		Bwt. 1	Bwt. 5
Adn(all)	133	39%	83%
	2/3 m/d	68	96
(74)	3 comm	54	90
	Oth	62	88
	Tot	57	89
Adn(Sug.	133	-	
modif.)	2/3 m/d	68%	96%
	3 comm	68	96
(51)	Oth	62	88
	Tot	65	92
Adn(N6	133	28%	78%
modif.)	2/3 m/d	56	94
(50)	3 comm.	42	86
	Oth	43	79
	Tot	42	84
Pur	133	41%	82%
	2/3 m/d	68	96
(74)	3 comm	55	89
	Oth	59	89
	Tot	57	89

(a)

		Bwt. 1	Bwt. 5
Pyr	133	78%	89%
	2/3 m/d	67	78
(26)	3 comm	72	83
	Oth	25	50
	Tot	58	73
All(excl.	133	50%	84%
C-glyc.)	2/3 m/d	70	92
	3 comm	61	87
(105)	Oth	50	78
	Tot	57	85
C-glyc	133	63%	75%
	2/3 m/d	37	37
(20)	3 comm	50	56
	Oth	0	0
	Tot	40	45
C-glyc	133	62%	87%
	2/3 m/d	62	62
(20)	3 comm	62	75
	Oth	75	100
	Tot	65	80

(b)

Table 4.2: Base weight determination. Correct value ranked first ("1") and amongst the top five candidates ("5") for seven structural categories, subdivided according to sugar type (see caption to table 4.1). Base weight determined by procedures of program NUCL dealing with (a) C-N bonded nucleosides, and (b) C-C bonded carbon glycosides for this category alone.

4.3.2 Presentation of Results Percentage success of the nucleoside weight determination routine is summarised in table 4.2 and graphed in figures 4.3 and 4.4, which are of similar composition to figures 4.1 and 4.2. As only one method was applied to all categories the graphs have been compressed slightly from the earlier figures, with the three sugar types plotted in figures 4.3(a) and 4.4(a) and total nucleosides, except for carbon glycosides, plotted alone in figures 4.3(b) and 4.4(b). The results of the carbon glycoside routine, specific to this class, have been indicated as single points on the graphs of figures 4.3 and 4.4. Again the full results from which these graphs have been constructed are presented in appendix III.

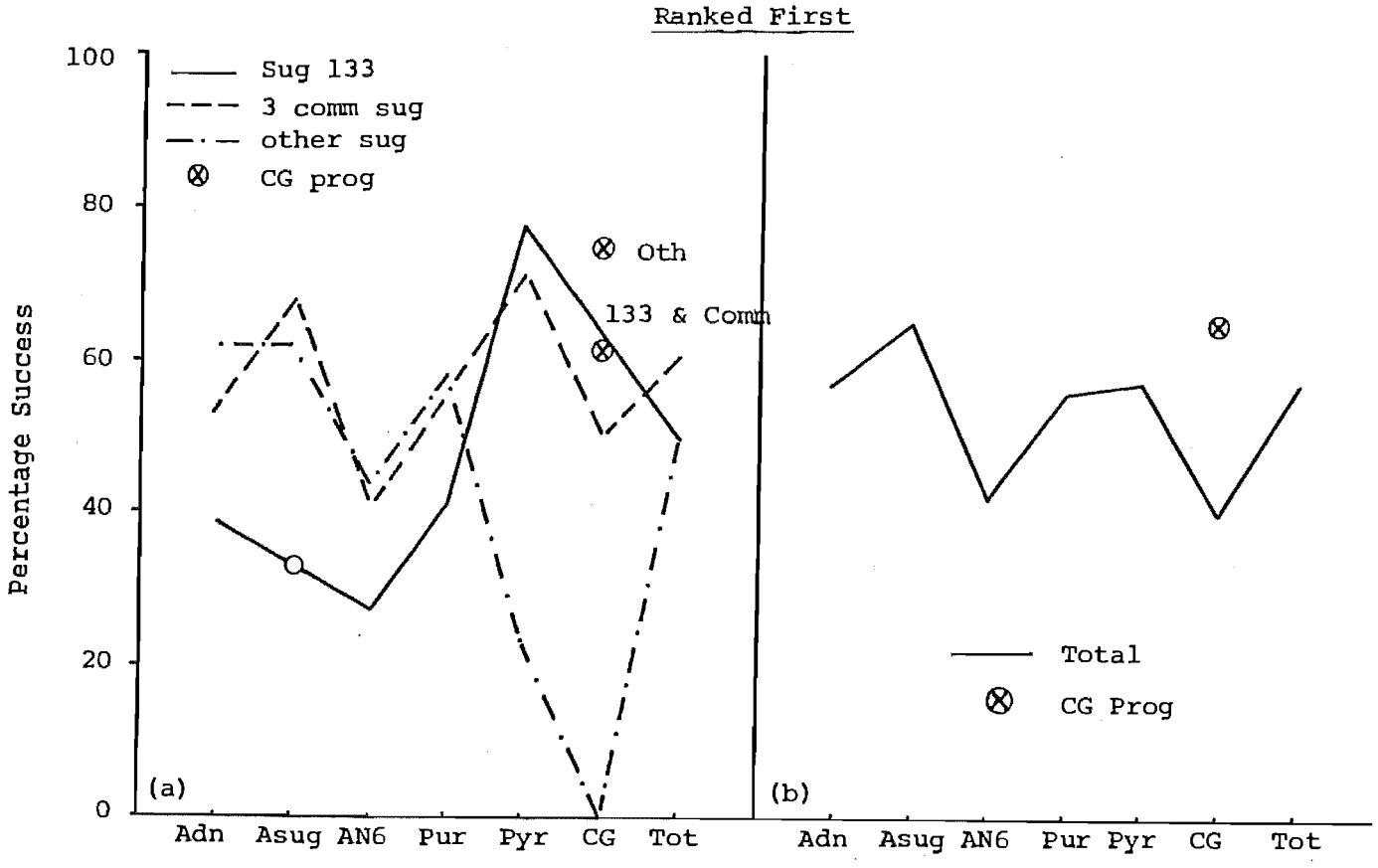


Figure 4.3: Base weight determination: correct candidate ranked first. Graphs plotted for (a) three sugar types (subsection 4.2.2), and (b) total nucleosides excluding carbon glycosides. Results of carbon glycoside (CG) program on this class alone plotted as single points.

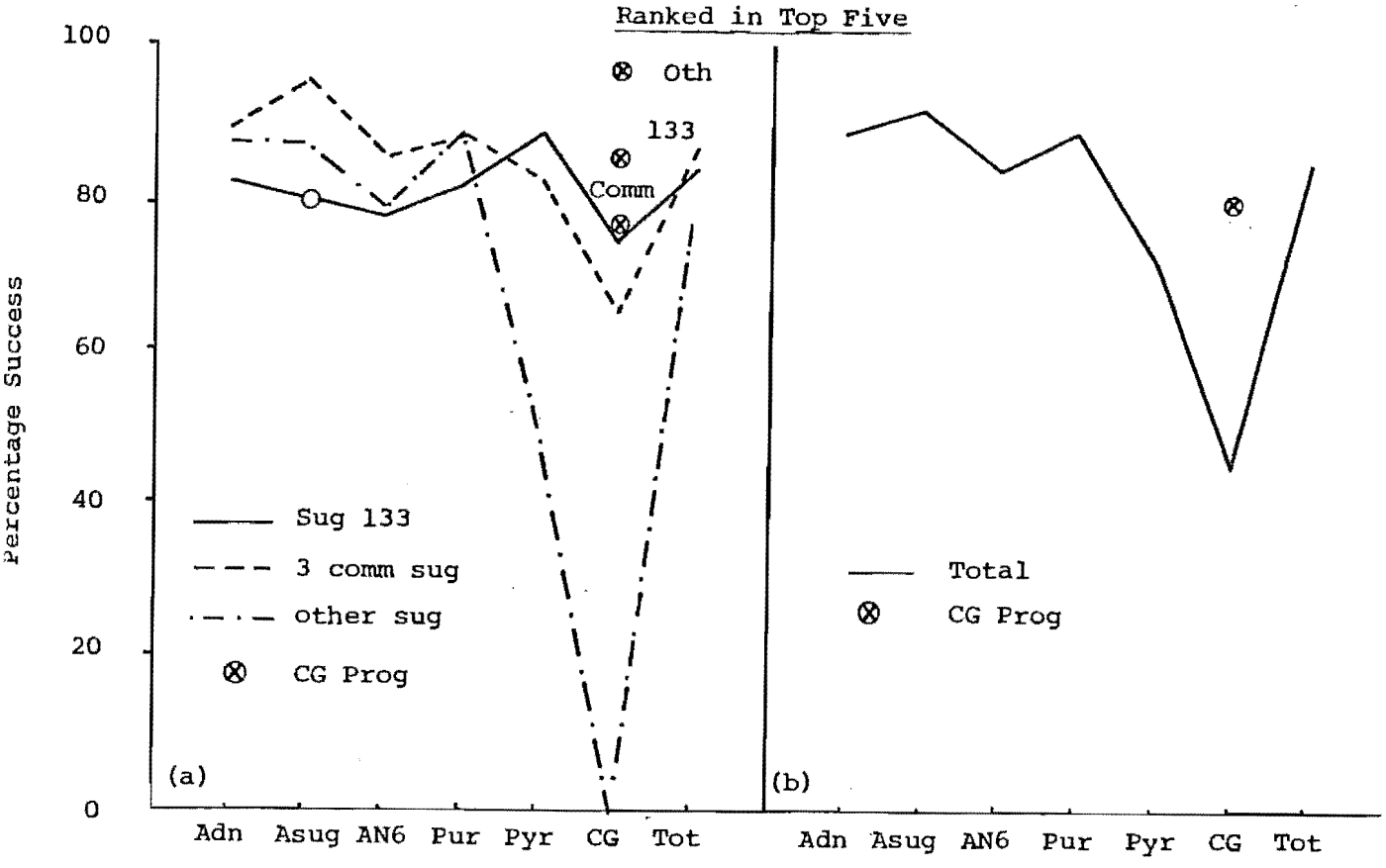


Figure 4.4: Base weight determination: correct candidate ranked amongst the first five. Graphs plotted for (a) three sugar types (subsection 4.2.2), and (b) total nucleosides excluding carbon glycosides. Results of carbon glycoside (CG) program on this class alone plotted as single points.

4.3.3 Results and Discussion The first part of this subsection applies only to C-N bonded nucleosides as opposed to carbon glycosides which are discussed separately. The correct value of the base mass B was ranked first in 57% of the total number of cases (figure 4.3(b)) although in 85% it was ranked amongst the top five candidates (figure 4.4(b)).

The criterion of being first ranked candidate leads to wide variations between the three sugar types (figure 4.3(a)) whereas relaxation

of this requirement to merely inclusion in the top five has, with one or two exceptions, a marked smoothing effect as well as an obvious increase in performance (figure 4.4(a)). If the behaviour of the program on carbon glycosides (the penultimate data point in each of the graphs) is excluded from consideration, then both of the effects are enhanced. The behaviour of this program on carbon glycosides is in fact rather misleading, given that it was not designed for such compounds which are dealt with by a separate routine. As could perhaps be expected the program performed slightly less efficaciously on compounds with non standard sugars than on the more common types. This effect is particularly marked for pyrimidines but again is minimised by requiring only inclusion in the top five candidates (figure 4.4(a)).

The carbon glycoside routine performed well on the spectra for which it was designed, and had minimal effect on any others. In 65% of the carbon glycoside spectra the correct base weight was ranked first, and in 80% it was included in the top five candidates. Sugar type had relatively little effect (figure 4.4(a)), and overall performance (the isolated point of figure 4.4(b)) was comparable with that of the nucleoside routine on the other spectra. It should in conclusion be noted that while the selection of carbon glycosides (appendix I) was as general and as representative as it could be made, the small sample size (20 spectra) must render these results slightly less reliable than those for the larger set. This factor is of course less critical for a heuristic approach than for the statistical and pattern recognition studies of the later chapters, when consideration of this class in isolation was not attempted for this reason.

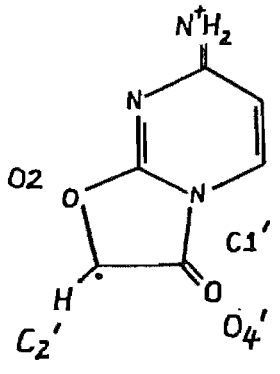
4.4 Identification of Nature of Base

An attempt was made to identify the base part of an unknown nucleoside as one of the four common types viz. cytosine, uracil, guanine or adenine. Characteristic losses from B+1 and characteristic ions at given mass values were used in this approach, which proved however almost completely fruitless.

4.4.1 Method The approach entailed a search of an unknown spectrum for the base derived ions described in subsection 3.2.2 and illustrated in schemes 3.5 - 3.8. By the inclusion of both losses and the constant m/z value ions formed by such losses from the standard unsubstituted bases, it was hoped to achieve recognition of most simply substituted bases. Underlying the method were the assumptions that (a) some fragments from B+1 would carry away with them the substituents, leaving ions at the same masses as those from the unmodified bases, and that (b) other fragments would not involve the substituents and therefore the losses would be of equal mass values to those from the unsubstituted compounds. These features are well illustrated by scheme 3.7 for 1- and 7- alkyl guanines. A further consideration was that the lower regions of many spectra, certainly below 69 amu and often below a higher value, are generally not reported and consequently ions utilised by the program should not fall below some such cut-off value.

This part of program NUCL relied heavily upon a knowledge of the base mass determined as reported in section 4.3, and base candidates from this earlier part were used as input to the base type procedure. It was hoped that this latter, by testing each candidate for consistency with the observed spectrum, might also have served as a check upon the correctness of the base weight.

The losses utilised for each type of nucleoside were as follows. For uridine (scheme 3.5) only one loss was considered: that of HNCO (43 amu) giving a characteristic ion from the unsubstituted uracil B+1 (112 amu) at 69 amu. All other common losses gave for at least the unsubstituted case ions below 69 amu. Cytidines (scheme 3.6) were characterised by losses of NH_2 (16 amu), CO (28 amu) and NCO (42 amu) giving characteristic ions derived from B+1 (111 amu) for the unsubstituted case at 95, 83 and 69 amu. In addition an ion unique to cytidines at B+41 of structure [267]:



B+41
(151 amu)

was also incorporated; the value derived from cytidine itself being 151 amu. Guanosines (scheme 3.7) were identified by losses of NH_2 (16 amu), NH_3 (17 amu), HNCN (41 amu), H_2NCN (42 amu), HNCO (44 amu), H_2NCN + CO (70 amu), and H_2NCO + CO (72 amu), yielding constant mass ions at 135, 134, 110, 109, 107, 81 and 79 amu from the unmodified species (151 amu). The programming for adenosines (scheme 3.8) involved losses of NH_2 (16 amu), HCN (27 amu), and HCN + HCN (54 amu), yielding ions at 119, 108, and 81 amu. Also a loss of CH_3 (15 amu) and corresponding ion at 120 amu was included as this is common in the spectra of N-6 substituted adenosines [268].

4.4.2 Results Application of this method to the present data base returned essentially meaningless results. It was not possible to distinguish base types by such an algorithm; such structural indications as were output were in many cases incorrect and misleading. Contributing causes may partly arise from the spectra used for testing as well as from the algorithm itself, and probably include the following:

(1) poor quality of some spectra in the testing set as evidenced by the fact that low mass range ions were often not reported,

(2) conversely, the presence of ions at nearly every low mass range position in other spectra, arising perhaps from background noise or low molecular weight impurities,

(3) variability of the losses from different compounds,

(4) an uneven weighting for the four classes, from one loss for cytidines to seven for guanosines, and

(5) gross overlap of losses e.g. 16 amu, and ions e.g. 69 amu, between base types.

4.5 Programmatic Details

Program NUCL was written in ALGOL and run in batch mode on the Burroughs B6718 at the University of Canterbury (section 5.4). The external program MOLION was supplied in FORTRAN and bound in in this form as a procedure of program NUCL. Nucleoside spectra were input to the program from punched cards, although paper tape or magnetic tape storage could also be accessed. Execution time varied slightly with the size of the spectrum but was typically 2-4 sec CPU.

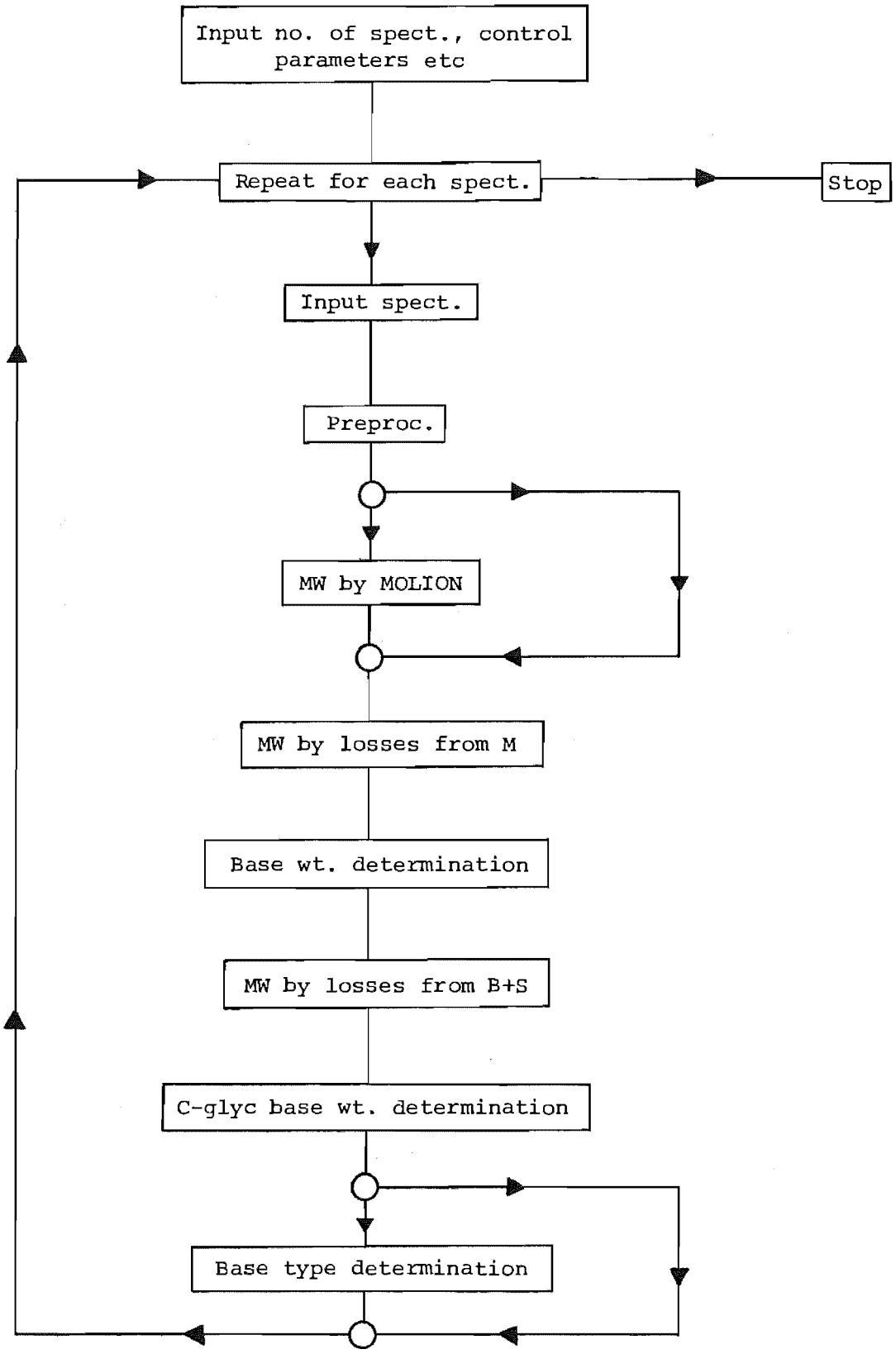


Figure 4.5: Schematic structure of program NUCL. Options within the program are indicated by alternative pathways.

Because of the small core usage and the limited number of spectra to which the program was initially applied, minimal attempt at optimisation was made. Approximately 500 lines of output were generated for each spectrum, much of which was an enumeration of possible structures and significances for the major ions. This was for use in later manual interpretation. A full listing of NUCL excluding MOLION is reproduced in appendix II, together with a sample output. Figure 4.5 is a broad overview of the program's structure.

Chapter 5

METHODS AND DATA

The heuristic (chapter 4) and pattern recognition (chapters 6-9) approaches are completely different both in method and, very often, in aims. This chapter distinguishes the two and reviews the concepts necessary for the later work (sections 5.1 and 5.3). The data base used throughout is also described (section 5.2).

5.1 Pattern Recognition Methodology

5.1.1 Basic Concepts A heuristic program encodes sets of fragmentations related in some known fashion to certain aspects of structure. Thus the aims of such a program are to deduce the structural features traditionally obtained from mass spectrometry such as molecular weight, etc. A pattern recognition approach, on the other hand, is not bound by known relationships of spectrum and structure. Thus structural features which are presumably related to the spectrum but by some as yet completely unknown relationship, features such as carbon number, oxygen number, certain very general patterns of substitution, etc (subsection 5.2.3) can hopefully be elucidated. As well, the more traditional structural questions can also be studied.

Unlike a heuristic program, which can be derived more or less from first principles, a pattern recognition program needs to be "trained" on some set of spectra the assignment of which, for the structural question under consideration, is known. Then, to evaluate the efficacy of this training, the pattern recognition classifier so developed must be applied to an "unknown" set of spectra, i.e. one which was not used in the training process, and the classifications achieved on the unknowns compared with their actual structure. This is termed prediction. The term "unknown" is slightly misleading, in so far as the structure of such spectra must be known to the programmer otherwise evaluation of the prediction classifications would be impossible. Such spectra are "unknown" to the programmatically developed classifier.

Fundamental to any pattern recognition study is the representation of a mass spectrum, or of selected mass positions from it, as an n-

dimensional pattern vector in an n-dimensional hyperspace, one dimension for each mass position. By selection of the mass positions used for each structural category it is hoped to make the spectra belonging to that category "cluster" in one part of the hyperspace, and those not belonging to that category "cluster" in another part. Hence an (n-1)-dimensional hypersurface can be constructed to separate the two classes. In the examples considered in this work this surface is linear, although non linear algorithms are possible. The construction of the decision surface is the function of the training process. This may simply involve calculation of some constant property of the two classes, such as the two n-dimensional means. Then the decision surface is merely the perpendicular bisector of the two means; spectra lying on one side of the surface are classified as class members and spectra lying on the other are classified as class non members. This is the approach adopted in chapter 8. Various refinements to this basic concept may be introduced, such as a region of non classification on either side of the decision surface, or the weighting of certain mass values.

If the mass spectra exhibit a fair degree of scatter in the hyperspace some may lie on the wrong side of the decision surface, even in the training set. Thus the surface may not classify 100% correctly the spectra from which it was formed. This leads to the concept of "recognition" i.e. the success of the classifier on the training set, as opposed to "prediction" on a set of unknowns.

An alternative approach is not to calculate some feature characteristic of the training set as a whole, such as the two class means, but to treat each spectrum individually as in the k-nearest neighbour method (chapter 9). In this method there is no training process as such, and assignment of unknowns is simply according to their distance from the individual members of the training set and the class membership or non membership of these.

One problem is the number of dimensions to choose for a given analysis i.e. the number of mass positions taken as a subset of the full mass spectral range. This dimensionality has a significant bearing upon the classification. In the limit as the dimensionality approaches the full number of mass values, say m/z 1-755 for the spectra here, the spectrum becomes exactly described by the associated pattern vector. If any kind of parameter fitting is attempted then the number

of cases, i.e. of spectra, must significantly exceed the dimensionality. It is now commonly accepted [269] that this excess should be at least by a factor of 3, and this has been adhered to in the present work. If no parameter fitting is involved, as in simple calculation of the means, then of course this restriction does not apply.

5.1.2 Terminology There are a number of terms commonly used in pattern recognition work which may require explanation. In their attempt at standardisation Harper et al. [271] have published a list of definitions, and their usage has generally been adhered to here. A number of terms can be used somewhat ambiguously as they have a strict pattern recognition definition which differs slightly from their more general usage, and while the exact meaning should be abundantly clear from context the opportunity is taken here to point out some possible sources of confusion.

binary: may refer to either binary data i.e. the peak/no peak form (subsection 5.2.2), or to a binary classification problem i.e. the classification of a test set into the two disjoint categories of members and non members.

case: an individual spectrum, also termed (in slightly different contexts) a pattern vector or a point in n-dimensional hyperspace. Can also be used to refer to a structural feature classification.

class member: a spectrum of a compound possessing a certain structural feature.

class non member: a spectrum of a compound lacking a certain structural feature.

classification, categorisation: used synonymously to refer to division of a test set of spectra into two disjoint subsets, class members and non members.

composition: may refer to either an elemental composition analysis as a type of structural feature e.g. total carbon number $\geqslant 12$, or to the composition of a test set of spectra i.e. how many spectra of which types it contains.

dimensionality: the number of mass positions chosen for a given classification problem.

error correction feedback: (chapter 7) used synonymously with the term learning machine approach.

feature: in a strict pattern recognition sense this refers to a particular component of a pattern vector i.e. to a mass position. Is also used in the sense of structural feature i.e. an aspect of molecular structure, and in the more general sense of prominent characteristic e.g. a feature of the data or of the analysis under consideration.

population: the set of all possible nucleoside spectra, whether or not they have been used in this work or even as yet recorded. It is assumed that the subset of 125 spectra (subsection 5.2.1) of this population chosen here is representative of the overall population.

Tr76, Pr20, Pr49: defined in subsection 5.2.1.

5.1.3 Similar Pattern Recognition Studies A large bulk of previous work is reviewed in chapter 2, and to give a concise summary of the most closely related pattern recognition work the most relevant publications are listed here. These constitute in large part the foundation of chapters 5-9.

(a) The book of Jurs and Isenhour [35] on pattern recognition in chemistry, a survey of the field.

(b) A recently published collection of symposium papers edited by Kowalski [270] on pattern recognition and other computer techniques in chemistry. The appellation and associated concept of "chemometrics" advanced here is an attempt at standardisation and unification in the field.

Four separate studies very similar in aims and methods to chapters 6-9.

(c) A paper of Jurs and Isenhour [143] describing the basic methods used in the later chapters.

(d) Two papers by Rotter et al. [145,146] on pattern recognition of steroid mass spectra, using the distance from the mean approach (cf. chapter 8) and including [146] a description of the classification evaluation methods of section 5.3.

(e) A paper of Wilkins et al. [141] expanding and refining the evaluation methods proposed by Rotter et al. above.

5.2 The Data Base

5.2.1 Composition of Test Sets

Low resolution 70eV electron impact mass spectra of underivatised nucleosides were taken from the chemical literature of the period 1962-78. A small number of spectra were also obtained on the AEI MS-902 mass spectrometer in the Chemistry Department of the University of Canterbury, the instrumental system of which has been described by Wright et al. [113]. These were both of compounds whose spectra had not previously been reported and also, for comparison purposes, of compounds whose spectra had been obtained by other researchers. Many published spectra were rejected from the file either on the grounds of excessive structural variation, such as derivatisation, or of too few peaks in the spectrum. Most guanosines, for example, fell into this latter category. A complete list of the 125 compounds whose spectra were used in this work is given in appendix I. The elemental compositions of the 125 nucleosides lay in the range $C_{9-32}H_{11-46}O_{2-8}N_{1-8}S_{0-1}F_{0-1}Cl_{0-1}Br_{0-1}$. The molecular weights lay between 211 and 755 amu.

Initially a file of 96 suitable spectra was obtained and the early pattern recognition studies performed on these. The spectra were randomly divided into a training set of 76 (Tr76) and a prediction set of 20 (Pr20). It was desired to maintain, for each category on which classification was attempted, approximately the same ratio of class members to non members in both the training and the prediction sets. Thus the number of class members that each prediction set should contain was calculated, and prediction set selection was performed simply by taking the last class members of the file of 96 until this number was reached. Then the last class non members were extracted until the prediction set had been made up to 20, and the remaining 76 spectra were used as the training set. Thus the training sets for each of the categories are not identical but differ in a few spectra.

Later, a further check of the literature was made and 29 more spectra obtained. These were added to the first prediction set of 20 spectra to form a second prediction set of 49 (Pr49), and the weight vectors developed on the original training set were tested on this augmented prediction set. The new spectra were added to form Pr49 in approximately the same ratio of class members to non members as exhibited by the training

set, thus indicating that the initial selection was a fair approximation to the distribution of the population. For a few categories however this did not hold and so the class composition of Pr49 differs somewhat from that of Tr76 and Pr20. This can be seen from, for example, table 6.4. The value of reporting two different prediction sets, as has been done consistently throughout chapters 6-9, lies in the resultant reinforcement of the independence of the training and the prediction processes. All the pattern recognition studies surveyed in subsection 1.3.3 use a single prediction set, and more widespread acceptance of their results could perhaps be enhanced by testing on a second prediction set drawn largely from fresh sources. Considerations of this nature are especially important with a small data base such as that used here.

5.2.2 Spectral Pre-processing The data obtained as described above was stored on magnetic tape and accessed for the various analyses by a set of FORTRAN input routines. Peaks below m/z 100 were removed from the spectra as in many of the references this range has been only scantily reported. Each spectrum was normalised to 100% of the base peak, and peaks of less than 1% relative intensity were removed. Two forms of intensity pre-processing were applied. The first was a reduction of the spectra to binary (peak/no peak) form, in which representation a peak of any intensity $\geqslant$ 1% is represented by "1" and absence of a peak is denoted by "0". This form has been shown [272] to retain a high information content whilst greatly reducing computation. The second was to take logarithmic spectra by the transformation

$$\log_{10}(\text{int} + 1) \qquad (5.2.1)$$

thereby reducing the dynamic range of the data from 0 - 100 to 0 - 2. In accordance with a suggestion of Kowalski and Bender [223] this representation was further transformed by normalisation to zero mean and unit standard deviation, a process termed by them "autoscaling". In general, as will be seen from the ensuing chapters, binary data retained only slightly less information content than the autoscaled logarithmic form, although it sometimes suffered from greater convergence difficulties. In very large applications this advantage of the logarithmic form may be offset by the greater computation involved.

5.2.3 Structural Categories The pattern recognition analyses of chapters 6-9 were conducted on twenty-one structural features of the base fragment, the sugar portion, and the nucleoside as a whole. These ranged from elemental composition to substitution pattern and nature of the base or sugar. These categories, defined in the order in which they are listed in the results tables of chapters 6-9, are as follows.

CT11, CT12, CT15: total carbon number, i.e. composition of the nucleoside as a whole, $\geqslant$ 11, 12 and 15 respectively.

OT5, OT6: Total oxygen number $\geqslant$ 5 and 6.

C6, C7, C8, C10: carbon number in the base alone $\geqslant$ 6, 7, 8 and 10.

O1, O2: oxygen number in the base alone $\geqslant$ 1 (oxygen presence/absence) and 2.

N4, N5: nitrogen number in the base alone $\geqslant$ 4 and 5. These two categories very nearly exactly correspond also to total nitrogen number $\geqslant$ 4 and 5.

NC6: nitrogen functionality at C-6 in purine heterocycles and at C-4 in pyrimidine heterocycles. This common substitution pattern is characteristic of adenosines, cytidines, etc.

OC2: oxygen functionality at C-2 in the base heterocycle. This is characteristic of cytidines, uridines, etc.

Pur, Pyr: base type purine and pyrimidine. These two are not complementary subsets of the data base as a number of nucleosides fall into neither category.

Adn: base type adenine, the most common single base (59.2% of the training set Tr76).

AN6: adenine type base substituted at the N-6 position, a form common to many naturally occurring nucleosides.

Asug: adenine type base with a modified sugar moiety i.e. one that is not D-ribose.

S133: sugar type D-ribose, of formula weight 133.

The structural features which could be examined were limited by the restriction that the two groups (class members and non members) be not grossly different in size. Varmuza, Rotter and Krenmayr suggest [146] a ratio within the range 70:30 and as can be seen from the p(1) column of, for example, table 6.2, this was adhered to in all but two cases. The a priori class membership probabilities (p(1)) lay

between 0.303 and 0.697 except for OC2 (0.250) and Pyr (0.211), two categories whose major importance in the chemistry of nucleosides prompted their inclusion notwithstanding. A restriction of this nature is especially important with small training sets; however it means that many chemically interesting structural categories such as carbon-glycoside, cytidine, uridine, etc could not be treated.

5.3 Classification Evaluation

5.3.1 Evaluation Requirements Once a classifier has been applied to a test set of data, either for recognition purposes on the set on which it was trained, or for prediction on an unknown set, the results obtained must be evaluated. This evaluation needs to fulfill several functions:

(a) it must render the results comparable with those from other classifiers on the same test set,

(b) it must render the results comparable with those for the same classifier on other, differently constituted, test sets, and

(c) it must convey an idea of the "goodness" of classification obtained.

Any such evaluation is necessarily a function of three independent factors:

(i) the a priori class probability i.e. the composition of the test set in terms of the proportion of class members p(1) and of non members p(2) = 1-p(1),

(ii) the success rate P_1 on class members, termed the class conditional probability and defined as the number of members correctly classified divided by the total number of members, and

(iii) the class conditional probability P_2 for class non members, defined analogously to P_1.

A possible fourth factor, the actual size N of the test set, is not generally taken into account.

The simplest, most intuitively obvious measure of classification is the overall success probability, P_{tot}, which is the number of cases correctly classified divided by the total number of cases. It will be obvious that

$$P_{tot} = p(1)P_1 + p(2)P_2 \quad . \tag{5.3.1}$$

The measure suffers from the disadvantage of not fulfilling criterion (b) above i.e. the composition of the test set can greatly affect the apparent "goodness" of classification. The following example may make this clear. Consider a test set composed of 99% class members and only 1% non members (a priori probability of class membership = 0.99). A classification success of P_{tot} = 70% by some classifier is intuitively much less acceptable than an equal classification success of P_{tot} = 70% on some other test set composed of, say, 50% members and 50% non members. Various proposals have been advanced to minimise this dependence on test set composition; these are outlined in this section and are applied to the classifiers of chapters 6,7,8 and 9.

One of the most obvious remedies is by comparison with blanket assignment of all cases in the test set to the more populous class. Thus in the example above the classification of P_{tot} = 70% on the 0.99 set would yield an improvement over classification by assignment to the more populous category of -19%, a very poor result indeed, whereas on the 0.50 set improvement would be +20%. This however is a coarse measure and a more sophisticated approach has been taken with the borrowing of certain concepts from information theory as described in the next subsection.

5.3.2 Information Theory This step was first taken in a chemical context by Rotter and Varmuza [273] in 1975 and has since been extended by Wilkins et al. [141]. A brief description of the information theory foundation of the measures will be given, which although neither derivationally complete nor mathematically rigorous may yet provide some understanding of the approach.

Information theory is concerned with the communication of information, and the following concepts have been defined. Let some event E_i occur with probability $P(E_i)$. On being told of its occurrence one can claim to have received an amount of information $I(E_i)$ defined as

$$I(E_i) = \log_2 \frac{1}{P(E_i)} \qquad (5.3.2)$$

bits. For example, if E_1 represents occurrence of a head on the tossing of a coin then $P(E_1) = 0.5$ and $I(E_1) = 1$ bit, i.e. "one bit is the amount of information one obtains when one of two possible equally likely

alternatives is specified" [274]. The probability that one will obtain an amount $I(E_i)$ of information is just $P(E_i)$, and so if there are q possible events then the average amount in bits of information obtained per event is

$$H(E) = \sum_{i=1}^{q} P(E_i) \log_2 \frac{1}{P(E_i)} \qquad (5.3.3)$$

H(E) is termed the entropy of the source and can alternatively be regarded as the average amount of uncertainty which the observer has before his inspection of the output of the source. In the coin tossing example above H(E) is of course one bit.

Thus it can readily be seen that for two classes A and B, if the a priori probability of membership in A is p(1) and in B is p(2) = 1-p(1), then the a priori entropy or uncertainty H(A) regarding class membership is

$$H(A) = p(1) \log_2 \frac{1}{p(1)} + p(2) \log_2 \frac{1}{p(2)} \qquad (5.3.4)$$

This is the uncertainty before application of the classifier; before arriving at the uncertainty after application some further probabilities must be defined. In the following the original terminology of Rotter and Varmuza [273], adopted also by Wilkins et al. [141], has been adhered to.

"1" and "2" denote class membership and non membership respectively in a binary classification problem. "j"(ja) and "n" (nein) denote assignment by the classifier i.e. "j" refers to assignment as a class member and "n" to that as a class non member. The following probabilities are defined. $p(1|j)$ and $p(2|n)$ are the a posteriori probabilities of membership in classes 1 and 2 following application of the classifier i.e. $p(1|j)$ is the probability that a case actually belongs to class 1 given that the classifier says it does [141], and similarly $p(2|n)$ is the probability that a case belongs to class 2 (non members) given that it has been classified as such. These two may be of some value in determining efficacy of classification and accordingly have been tabulated for the classifications reported in chapters 6-9 along with the other measures. $p(1|n)$ and $p(2|j)$ are the analogous probabilities for incorrect assignment. p(j) and p(n) are the probabilities that a classifier will assign a case to class 1 and to class 2 respectively. Obviously

$$p(1|j) + p(2|j) = 1 \tag{5.3.5}$$

$$p(1|n) + p(2|n) = 1 \tag{5.3.6}$$

$$p(j) + p(n) = 1 \quad . \tag{5.3.7}$$

Thus we can now define the residual uncertainty $H(A|B)$ after application of the classifier

$$H(A|B) = p(j)H(A|j) + p(n)H(A|n) \tag{5.3.8}$$

where

$$H(A|j) = p(1|j) \log_2 \frac{1}{p(1|j)} + p(2|j) \log_2 \frac{1}{p(2|j)} \tag{5.3.9}$$

and

$$H(A|n) = p(1|n) \log_2 \frac{1}{p(1|n)} + p(2|n) \log_2 \frac{1}{p(2|n)} \quad . \tag{5.3.10}$$

Hence we arrive at another measure of the efficacy of a classifier, the information gain $I(A,B)$ defined as the decrease in uncertainty engendered by application of the classifier

$$I(A,B) = H(A) - H(A|B) \quad . \tag{5.3.11}$$

An example may illustrate the meaning of this measure. If, in a two class problem such as the tossing of a coin, the outcome is completely random, i.e. the a priori probability $p(1)$ is 0.5, then the initial uncertainty $H(A)$ as defined in equation (5.3.4) is a maximum, $H(A) = 1$ bit. If we are told that the outcome is a head, so that the residual uncertainty is zero, the information gain is the maximum possible, $I(A,B) = 1-0 = 1$ bit. If however for some reason we are only 90% sure that the outcome is a head, then

$$p(j) = 0.9,\ p(n) = 0.1,\ p(1|j) = p(2|n) = 0.9,$$

$$p(2|j) = p(1|n) = 0.1$$

and so

$$H(A|j) = H(A|n) = 0.9 \log_2 \frac{1}{0.9} + 0.1 \log_2 \frac{1}{0.1}$$

$$= 0.468 \quad .$$

Consequently

$$H(A|B) = 0.9 \times 0.468 + 0.1 \times 0.468$$

$$= 0.468$$

and thus

$$I(A,B) = 1-0.468 = 0.532 \quad .$$

Thus an uncertainty in the final outcome of only 10% results in an information gain of just over half a bit.

It is obvious that the maximum possible information gain as defined in equation (5.3.11) will not be the same for all test sets, but will depend upon H(A) which in turn depends upon p(1) and p(2) the test set composition. Two nearly equivalent ways of overcoming this have been proposed. The first, by Rotter and Varmuza [273], was to reduce each value of I(A,B) to that which would be obtained on a test set composed of equal numbers of members and non members. This measure they termed I_{max} since it can readily be shown (cf. the symmetry of figure 5.1) that the information gain for such a test set is a maximum. Wilkins et al. [141] in the other approach proposed a figure of merit M, defined as the information gain relative to the maximum possible information gain imposed by the composition of the test set

$$M = \frac{I(A,B)}{H(A)} \quad . \tag{5.3.12}$$

This equation arises because the minimum possible residual uncertainty $H(A|B)$ after application of the classifier is of course zero, in which case by equation (5.3.11) I(A,B) reduces to H(A). In the example given after equation (5.3.11), if the outcome is exactly known H(A) = I(A,B) = 1 bit and so M = 1. For the example with only 90% certainty in the final outcome H(A) = 1 bit and I(A,B) = 0.532, so M = 0.532. If a biased coin had been used so that the a priori probability was, say, 0.7 instead of 0.5, then the initial uncertainty H(A) would have been 0.880 bit by equation (5.3.4) and so

$$M = \frac{0.880 - 0.468}{0.880} = 0.468 \quad .$$

Both these measures, I_{max} and M, have been tabulated in chapters 6-9.

5.3.3 Computational Formulae The performance evaluation measures P_1, P_2, P_{tot}, improvement over classification into the more populous category, $p(1|j)$, $p(2|n)$, I(A,B), I_{max}, and M have been calculated for each analysis by a small program CLASSIFMEASURE which

is reproduced in appendix II. This program is founded on the computational formulae given below for each of the measures. The classification tables reproduced in the later chapters were output directly by this program, as also were the histograms of classifier performance.

As mentioned in subsection 5.3.1 goodness of classification can be expressed in terms of four variables, size N of test set, number of members n_1, and numbers of members c_1 and of non members c_2 correctly classified. All the various probabilities and evaluation measures described above are function only of these four. It is also convenient to derive the number n_2 of class non members

$$n_2 = N-n_1 \quad . \tag{5.3.13}$$

Obviously the a priori class probability is

$$p(1) = \frac{n_1}{N} \tag{5.3.14}$$

and the class conditional probabilities are

$$P_1 = \frac{c_1}{n_1} \tag{5.3.15}$$

$$P_2 = \frac{c_2}{n_2} \tag{5.3.16}$$

whence the overall success rate

$$P_{tot} = \frac{c_1+c_2}{N} \quad . \tag{5.3.17}$$

The a posteriori class probabilities are given by

$$p(1|j) = \frac{c_1}{c_1+(n_2-c_2)} \tag{5.3.18}$$

for the probability that a case will belong to class 1 given that it has been classified as such, and

$$p(2|n) = \frac{c_2}{c_2+(n_1-c_1)} \quad . \tag{5.3.19}$$

(n_2-c_2) is the number of non members (class 2) misclassified as being in class 1, and (n_1-c_1) is the number of members misclassified into class 2. The analogous probabilities for incorrect assignment become

$$p(1|n) = \frac{n_1-c_1}{c_2+(n_1-c_1)} \tag{5.3.20}$$

$$p(2|j) = \frac{n_2-c_2}{c_1+(n_2-c_2)} \quad , \tag{5.3.21}$$

and the probabilities for classification into classes 1 and 2 are

$$p(j) = \frac{c_1+(n_2-c_2)}{N} \tag{5.3.22}$$

$$p(n) = \frac{c_2+(n_1-c_1)}{N} \quad . \tag{5.3.23}$$

The algorithm upon which program CLASSIFMEASURE is based is merely the incorporation of these expressions into equations (5.3.11) and (5.3.12).

5.3.4 Behaviour of the Functions One final point concerning the functions I(A,B) and M should be made. A previously unreported quirk of their behaviour is their misleadingly high values for cases of very poor classification. This arises because they are not monotonically increasing functions for all values of the class conditional probabilities P_1 and P_2. Figure 5.1 is a plot of the figure of merit M against P_1 for various P_2, at constant p(1) (the a priori class probability). As can be seen, unless one class is classified either completely correctly ($P_2 = 1.0$) or completely incorrectly ($P_2 = 0.0$) the function passes through a minimum at some value of P_1, $0 < P_1(\text{min}) < 1$. This value, as will be shown, depends solely on P_2. For example, as can been seen from figure 5.1 the curve $P_2 = 0.8$ passes through a minimum at $P_1 = 0.2$. Consequently, if one class is predicted with an accuracy of 80% ($P_2 = 0.8$) and the other with an accuracy of only 10% ($P_1 = 0.1$), the figure of merit value will be higher than if the latter class had been predicted with an accuracy of 20% ($P_1 = 0.2$), an obviously misleading result.

The expression (5.3.11) for I(A,B) can be expressed in terms of the class conditional probabilities P_1 and P_2

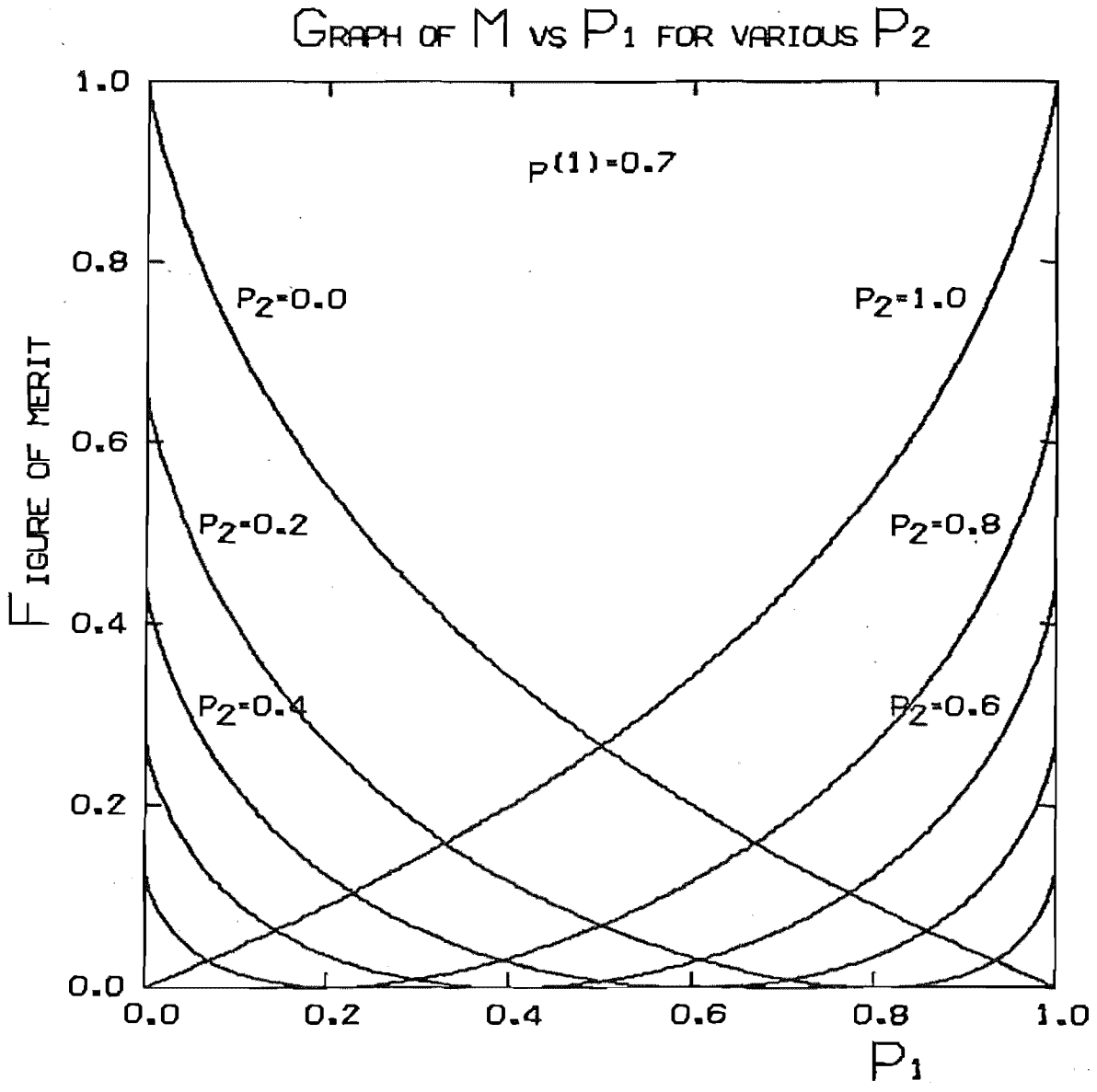


Figure 5.1: Theoretical curves for figure of merit M. Curves plotted as a function of class conditional probability P_1 for various P_2 and at constant test set composition (p(1) = 0.7).

$$I(A,B) = p(1)P_1 \log \frac{P_1}{p(1)P_1 + p(2)(1-P_2)} +$$

$$+ p(2)(1-P_2) \log \frac{1-P_2}{p(1)P_1 + p(2)(1-P_2)} +$$

$$+ p(1)(1-P_1) \log \frac{1-P_1}{p(1)(1-P_1) + p(2)P_2} +$$

$$+ p(2)P_2 \log \frac{P_2}{p(1)(1-P_1) + p(2)P_2} \quad . \qquad (5.3.24)$$

Differentiation with respect to P_1, holding P_2 constant, yields

$$\frac{d[I(A,B)]}{dP_1} = p(1) \log \left| \frac{P_1}{1-P_1} \cdot \frac{p(1)(1-P_1) + p(2)P_2}{p(1)P_1 + p(2)(1-P_2)} \right| \qquad (5.3.25)$$

which when equated to zero returns the value of P_1 at which I(A,B) or equivalently M is a minimum

$$P_1(\min) = 1-P_2 \quad . \qquad (5.3.26)$$

Logarithms are to base 2 and conversion from I(A,B) to figure of merit M is by multiplication by a factor 1/H(A) given in equation (5.3.4), constant for a given test set composition (p(1)). Both of these disappear on equating the differential to zero. This restriction on the valid range of I(A,B) and M is equivalent to ensuring that the sum of the two class conditional probabilities is at least unity

$$P_1 + P_2 \geq 1 \qquad (5.3.27)$$

and a check of this was made in each classification case.

5.4 Computing System

The pattern recognition analyses, like the program NUCL of chapter 4, were run in batch mode on the Burroughs B6718 at the University of Canterbury. The central processor unit (CPU) of this machine is a multi-programming data processor operating at 5 Mhz clock rate with vector hardware and an arithmetic unit operating at 10 Mhz clock rate. Core memory is 6 modules of 16K 48 bit words with a cycle time of 1.2 micro-seconds, and 2 modules of 64 K 48 bit words with a cycle time of 1.6

microseconds. The graphics device on which the histograms of chapters 6-9 were produced by program CLASSIFMEASURE (appendix II) is an 11 inch Calcomp X-Y plotter.

Programs written for this work are reproduced in appendix II together with sample outputs. CPU times for the pattern recognition analyses are listed in the methods sections of chapters 6-9.

Chapter 6

STATISTICAL DISCRIMINANT FUNCTION ANALYSIS

6.1 Introduction

A statistical linear discriminant function analysis was applied to the nucleoside data base described in subsection 5.2.1. The method (section 6.2) is based on an assumption of multivariate normality. It is considered to be statistically robust and application to a highly skew non Gaussian mass spectral intensity distribution provided a severe test of its generality. The method has previously been applied in other chemical fields such as the classification of petroleum pollutants by their IR spectra [77a], with excellent results.

6.2 Method

The statistically based discriminant function analysis used involved the computation of a set of binary linear classification functions to independently separate the twenty-one categories of nucleoside (subsection 5.2.3). The computations were performed by the program BMD 07M [275] for the multivariate analysis of variance. The program is based on established algorithms [276] and has been described in detail elsewhere [275, 77c, 278]. Consequently only a very abridged description is given here.

A classification function of the form

$$Y_k(\underline{x}_i) = \underline{c}_k\underline{x}_i \qquad (6.2.1)$$

to determine the membership of the (p+1)-dimensional ith pattern vector (i.e. spectrum) $\underline{x}_i$ in the kth category can be found such that $Y_k > 0$ for class members and $Y_k \leqslant 0$ for non members. $\underline{c}_k$ is a (p+1)-dimension coefficient matrix and the augmented variables x_{i0} are defined as unity. The program is capable of distinguishing more than one category simultaneously but this facility was not found to be of use. The following treatment refers to the classification of only two groups (k = 1,2) i.e. class members and non members. The method is strictly applicable only

to a population of multivariate normal distribution, an obviously invalid assumption for the highly skew one-tailed mass spectral intensity distribution at most m/z values. Notwithstanding this, however, the known statistical robustness of the approach [279], i.e. its insensitivity to deviations from normality, prompted its application to the present data both as a classification measure in its own right and as a further, severe test of the method's statistical robustness. This can be evaluated by comparison of the results obtained here with those from other, non statistical approaches such as distance from the mean (chapter 8) or k-nearest neighbour (chapter 9).

The within groups cross product matrix $\underline{W}$ is formed

$$\underline{W} = \underline{X}'\underline{X} - \bar{\underline{X}}'\underline{N}\,\bar{\underline{X}} \qquad (6.2.2)$$

where $\underline{X}$ is the p x n raw data matrix for p variables and n spectra, each row containing the data for a single spectrum, $\underline{N}$ is a 2 x 2 (in the case of only two groups) square diagonal matrix containing as the diagonal elements the sizes n_1 and n_2 of the two groups, and $\bar{\underline{X}}$ is the p x 2 matrix of variable means within each group. $\underline{X}'$ denotes the transpose of matrix $\underline{X}$. $\underline{W}$ is related to the within groups covariance matrix $\underline{V}$ by

$$\underline{V} = \frac{1}{n-2}\,\underline{W} \quad . \qquad (6.2.3)$$

The total cross product matrix $\underline{T}$ is also formed

$$\underline{T} = \underline{X}'\underline{X} - n\bar{\bar{\underline{X}}}'\bar{\bar{\underline{X}}} \qquad (6.2.4)$$

where $\bar{\bar{\underline{X}}}$ is the p x 1 row matrix containing the means for each variable over both groups. The classifiers are chosen so as to maximise the separation of the means of the two groups, consequently perfect classification, even on the training set, is not necessarily achieved. Thus the values of all the classification measures in table 6.2, excluding "improv most pop" and "I(A,B)" (subsection 6.3.2) are not necessarily unity. Increasing numbers of variables are included in the classification functions, these being added according to the F criterion below. In the final stage with all p variables included the

matrix $\underline{c}$ of classification function coefficients is

$$\underline{c} = (n-2)\bar{\underline{X}}\,\underline{W}^{-1} \tag{6.2.5}$$

and constant terms $\underline{c}_0$

$$\underline{c}_0 = -\frac{1}{2}\,\underline{c}\,\bar{\underline{X}} \quad . \tag{6.2.6}$$

Several related statistics are computed as measures of the efficacy of separation of the two groups. Variables are sequentially entered into the classification functions according to their F value i.e. the ratio of the between groups variance to the within groups variance

$$F_j = (n-p-1)\,\frac{w^{jj} - t^{jj}}{t^{jj}} \tag{6.2.7}$$

with degrees of freedom 1 and n-p-1, for the case where all variables have been entered. w^{jj} and t^{jj} are the diagonal elements of $\underline{W}^{-1}$ and $\underline{T}^{-1}$. The F ratio is a test of the null hypothesis i.e. the hypothesis that the groups are merely random samples drawn from the same population. The larger the value the greater is the probability that there are real differences between the groups. The F ratio is furthermore related by equation (6.2.9) to the separation of the two groups as measured by the square of the Mahalanobis distance function

$$D^2 = (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)\underline{W}^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)' \tag{6.2.8}$$

where $\bar{\underline{X}}_1$ is the row vector containing the variable means for group 1.

$$F = \frac{(n-p-1)n_1n_2}{pn(n-2)}\,D^2 \quad . \tag{6.2.9}$$

A further measure of the separation of the groups is Wilks' Λ [280], also known as the U statistic, which if all variables have been entered is given by

$$\Lambda = \frac{|\underline{W}|}{|\underline{T}|} \tag{6.2.10}$$

with degrees of freedom p, 1 and n-2. Comparison of equations (6.2.7) and (6.2.10) reveals a relationship between F and Λ which takes, in the case of two groups, the exact form

$$F = \frac{n-p-1}{p} \cdot \frac{1-\Lambda}{\Lambda} \qquad (6.2.11)$$

with degrees of freedom p and n-p-1.

The program BMD 07M is restricted to 25 variables in an analysis, so for consideration of what was potentially a 755 dimension problem (m/z 1-755) the following feature selection process was applied. First, attention was restricted to those 82 m/z positions which appeared most frequently in the training set. Secondly, for each category, the program was run in parallel on four disjoint sets of 21, 21, 20, and 20 mass positions drawn from the set of 82. Thirdly, those six mass values which provided the greatest discriminant power from each of the four runs were taken and a further analysis conducted using this new set of 24. The F ratio was used as the measure of discriminant power. Finally, the best twelve or fewer of these were taken and run separately. The rationale behind the choice of this final number was the empirical observation that performance at about this number of variables levelled off, and increased either only very slowly or not at all as the rest of the set of 24 was added in. This is well exemplified by the results for the category OC2 shown in table 6.1. Here is seen the increase in

No. of variables added	No. cases misclassified			
	Tr 76		Pr 20	
	Mem	Non mem	Mem	Non mem
4	1	2	1	1
6	1	1	0	1
* 8	0	1	2	1
10	0	1	2	1
12	0	1	2	1
14	0	1	2	1

24	0	1	1	0

Table 6.1: Performance of statistical linear discriminant analysis for oxygen functionality at C2 (OC2). Numbers of spectra misclassified are shown for increasing dimensionality of the discriminant function. Results are for the training set of 76 spectra, consisting of 19 members and 51 non members, and the first prediction set of 20 spectra with 5 class members.

recognition on the training set Tr76 as more variables are added in until after about eight, the number chosen for the final analysis, there is little further improvement. This trend applies also to the prediction sets although this was not of course a consideration in the selection of the dimensionality of the weight vector. More than twelve variables were used in only two cases, CT12 and NC6 (table 6.8), when performance with this number was unsatisfactory. For the comparisons of chapter 10, however, the slight overall improvement gained by inclusion of the full 24 dimensions has lead to these results being considered as the "best".

With 24 variables the CPU time for each analysis lay in the range 25-30 sec and with 8-14 variables it was approximately 10 sec.

6.3 Presentation of Results

6.3.1 Tables and Figures The results for the twenty-one structural categories of subsection 5.2.3 are presented in tables 6.2 - 6.4 for the training and the two prediction sets Tr76, Pr20, and Pr49, respectively. These results are presented as histograms in figure 6.1 for the evaluation measures P_{tot}, the information gain I(A,B), the figure of merit M, and I_{max}. These measures are described in subsection 6.3.2. Tables 6.2-6.4 and figure 6.1 give the results for the reduced sets of 8-14 mass positions, and the full 24-dimension analyses are presented in tables 6.5-6.7 and figure 6.2. Table 6.5 gives recognition on the training set Tr76 and tables 6.6 and 6.7 give prediction on the sets Pr20 and Pr49 respectively. Finally, the 8-14 most discriminatory mass positions are tabulated in table 6.8. Full 24-dimension weight vectors are reproduced in appendix III.

6.3.2 Evaluation Measures As stated in section 6.2 recognition on the training set will not always be 100%. Results for this set (table 6.2) and the two prediction sets (tables 6.3 and 6.4) are recorded in terms of the classification measures of section 5.3, where the column headings are as follows.

In the nomenclature adopted, "1" denotes class membership and "2" denotes class non membership. "Mem" is the number of class members for each analysis, this is presented as a proportion of the total set size

(76 for the training and 20 and 49 for the two prediction sets) under "p(1)". "P_1" and "P_2" are the proportions of members and non members classified correctly; these are termed class conditional probabilities by Wilkins et al. [141] and predictive abilities by Rotter and Varmuza [147]. "P_{tot}" is the total proportion of correctly classified spectra and as such is the simplest, most intuitive measure of performance. This measure however is very dependent upon test set composition and the measures of section 5.3 have been advanced to account for this. "Improv most pop" is the improvement of the proportion of correctly classified spectra over simple assignment of all the spectra to the most populous class i.e.

$$P_{tot} - \max(p(1),\ 1-p(1)). \qquad (6.3.1)$$

This has been suggested as a worthwhile measure of classifier performance in so far as blanket assignment to the more populous class (usually non members) should not give a better categorisation than a specially calculated weight vector. "p(1|j)" and "p(2|n)" (subsections 5.3.2 and 5.3.3) are the a posteriori probabilities of membership in classes 1 and 2 following application of the classifier i.e. the probabilities that a spectrum actually belongs to the class once it has been classified as such. "I(A,B)" is the information gain, the figure of merit "Fig mer" is the information gain relative to the maximum possible information gain imposed by the composition of the test set, and I_{max} is the information gain that would be obtained on a test set composed of equal numbers of members and non members (subsections 5.3.2 and 5.3.3). These last two measures, as might be expected, give very nearly the same ordering of classifiers.

All the above evaluation measures are presented here so that, as well as a comparison of the classifiers themselves, a comparison of the means of evaluating them may also be made. A more visual representation of the results is given in figure 6.1 where the data of the three tables has been presented in blocks of three histograms, one for each of the training and the two prediction sets. These histograms have been drawn for the measures P_{tot}, I(A,B), figure of merit and I_{max}. It can clearly be seen from this graphical depiction that the apparent "goodness" of classifier performance can vary greatly according to the means used to

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(1|J)	P(2|N)	I(A,B)	FIG MER	IMAX
1	CT11	47	.618	.957	.897	.934	0.316	.938	.929	.609	.635	.630
2	CT12	35	.461	.943	.999	.974	0.434	.999	.953	.842	.846	.840
3	CT15	29	.382	.931	.999	.974	0.355	.999	.959	.801	.835	.816
4	OT5	42	.553	.881	.999	.934	0.382	.999	.872	.708	.714	.726
5	OT6	25	.329	.840	.980	.934	0.263	.955	.926	.566	.619	.599
6	C6	35	.461	.943	.999	.974	0.434	.999	.953	.842	.846	.840
7	C7	31	.408	.968	.999	.987	0.395	.999	.978	.884	.906	.896
8	C8	25	.329	.999	.980	.987	0.316	.962	.999	.833	.912	.930
9	C10	23	.303	.913	.981	.961	0.263	.955	.963	.645	.729	.716
10	O1	39	.513	.974	.973	.974	0.461	.974	.973	.824	.824	.824
11	O2	24	.316	.958	.981	.974	0.289	.958	.981	.727	.808	.806
12	N4	53	.697	.981	.957	.974	0.276	.981	.957	.712	.805	.803
13	N5	43	.566	.953	.939	.947	0.382	.953	.939	.691	.699	.699
14	NC6	45	.592	.978	.935	.961	0.368	.957	.967	.736	.755	.749
15	OC2	19	.250	.999	.982	.987	0.237	.950	.999	.736	.907	.936
16	PUR	46	.605	.935	.900	.921	0.316	.935	.900	.572	.591	.591
17	PYR	16	.211	.999	.999	.999	0.211	.999	.999	.742	.999	.999
18	ADN	45	.592	.978	.903	.947	0.355	.936	.966	.681	.698	.690
19	AN6	33	.434	.939	.907	.921	0.355	.886	.951	.600	.607	.611
20	ASUG	32	.421	.969	.795	.868	0.289	.775	.972	.490	.499	.512
21	S133	25	.329	.840	.999	.947	0.276	.999	.927	.642	.702	.664
	AV.			.947	.958	.956	0.332	.958	.956	.709	.759	.756

Table 6.2: Statistical discriminant function analysis on Tr76 using 8-14 m/z values. For column headings see subsection 6.3.2.

Table 6.3: [Overleaf] Statistical discriminant function analysis on Pr20 using 8-14 m/z values. For column headings see subsection 6.3.2.

Table 6.4: [Overleaf] Statistical discriminant function analysis on Pr49 using 8-14 m/z values. For column headings see subsection 6.3.2.

Table 6.3

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(1|J)	P(2|N)	I(A,B)	FIG MER	IMAX
1	CT11	13	.650	.769	.857	.800	0.150	.909	.667	.279	.299	.309
2	CT12	9	.450	.889	.909	.900	0.350	.889	.909	.525	.528	.528
3	CT15	7	.350	.857	.923	.900	0.250	.857	.923	.473	.506	.505
4	OT5	11	.550	.818	.778	.800	0.250	.818	.778	.273	.275	.275
5	OT6	7	.350	.714	.923	.850	0.200	.833	.857	.325	.348	.341
6	C6	9	.450	.889	.999	.950	0.400	.999	.917	.744	.750	.739
7	C7	8	.400	.625	.917	.800	0.200	.833	.786	.251	.259	.254
8	C8	6	.300	.667	.929	.850	0.150	.800	.867	.276	.313	.305
9	C10	6	.300	.667	.929	.850	0.150	.800	.867	.276	.313	.305
10	O1	10	.500	.900	.800	.850	0.350	.818	.889	.397	.397	.397
11	O2	6	.300	.667	.929	.850	0.150	.800	.867	.276	.313	.305
12	N4	14	.700	.857	.833	.850	0.150	.923	.714	.325	.369	.379
13	N5	11	.550	.818	.999	.900	0.350	.999	.818	.617	.621	.634
14	NC6	12	.600	.833	.875	.850	0.250	.909	.778	.385	.397	.402
15	OC2	5	.250	.600	.933	.850	0.100	.750	.875	.214	.264	.256
16	PUR	12	.600	.833	.750	.800	0.200	.833	.750	.256	.264	.264
17	PYR	4	.200	.750	.875	.850	0.050	.600	.933	.214	.297	.311
18	ADN	13	.650	.923	.571	.800	0.150	.800	.800	.212	.227	.221
19	AN6	9	.450	.556	.636	.600	0.050	.556	.636	.027	.027	.027
20	ASUG	9	.450	.778	.636	.700	0.150	.636	.778	.129	.130	.131
21	S133	7	.350	.714	.923	.850	0.200	.833	.857	.325	.348	.341
	AV.			.768	.854	.831	0.202	.819	.822	.324	.345	.344

Table 6.4

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(1|J)	P(2|N)	I(A,B)	FIG MER	IMAX
1	CT11	34	.694	.735	.533	.673	-.020	.781	.471	.048	.054	.055
2	CT12	24	.490	.542	.840	.694	0.184	.765	.656	.120	.120	.120
3	CT15	14	.286	.643	.800	.755	0.041	.563	.848	.127	.147	.151
4	OT5	28	.571	.571	.571	.571	0.000	.640	.500	.014	.015	.015
5	OT6	14	.286	.571	.771	.714	0.000	.500	.818	.076	.088	.091
6	C6	22	.449	.591	.741	.673	0.122	.650	.690	.082	.083	.083
7	C7	15	.306	.600	.794	.735	0.041	.563	.818	.105	.118	.120
8	C8	13	.265	.538	.861	.776	0.041	.583	.838	.112	.134	.135
9	C10	11	.224	.636	.842	.796	0.020	.538	.889	.134	.175	.182
10	O1	27	.551	.593	.682	.633	0.082	.696	.577	.055	.055	.055
11	O2	11	.224	.727	.842	.816	0.041	.571	.914	.185	.241	.253
12	N4	37	.755	.865	.750	.837	0.082	.914	.643	.233	.290	.299
13	N5	30	.612	.767	.999	.857	0.245	.999	.731	.517	.537	.568
14	NC6	33	.673	.697	.563	.653	-.020	.767	.474	.044	.049	.050
15	OC2	12	.245	.833	.973	.939	0.184	.909	.947	.474	.590	.571
16	PUR	28	.571	.821	.571	.714	0.143	.719	.706	.122	.124	.123
17	PYR	10	.204	.700	.949	.898	0.102	.778	.925	.276	.378	.368
18	ADN	29	.592	.862	.550	.735	0.143	.735	.733	.141	.144	.143
19	AN6	17	.347	.647	.625	.633	-.020	.478	.769	.049	.053	.054
20	ASUG	19	.388	.789	.633	.694	0.082	.577	.826	.129	.134	.137
21	S133	15	.306	.667	.882	.816	0.122	.714	.857	.219	.247	.246
	AV.			.686	.751	.743	0.077	.688	.744	.155	.180	.182

evaluate it. For example, the simplest evaluation, P_{tot} (figure 6.1(a)), makes the results on the three sets appear very much better than do either I_{max} (figure 6.1(d)) or the figure of merit (figure 6.1(c)), the two most sophisticated. The three information gain criteria (figure 6.1 (b)-(d)) in fact return very similar results; this is due to reasonably well balanced test sets as $0.3 < p(1) < 0.7$ in all save two cases. This similarity is quantified in the next section, and for subsequent analyses in this work only figure of merit M and P_{tot} will be presented as histograms.

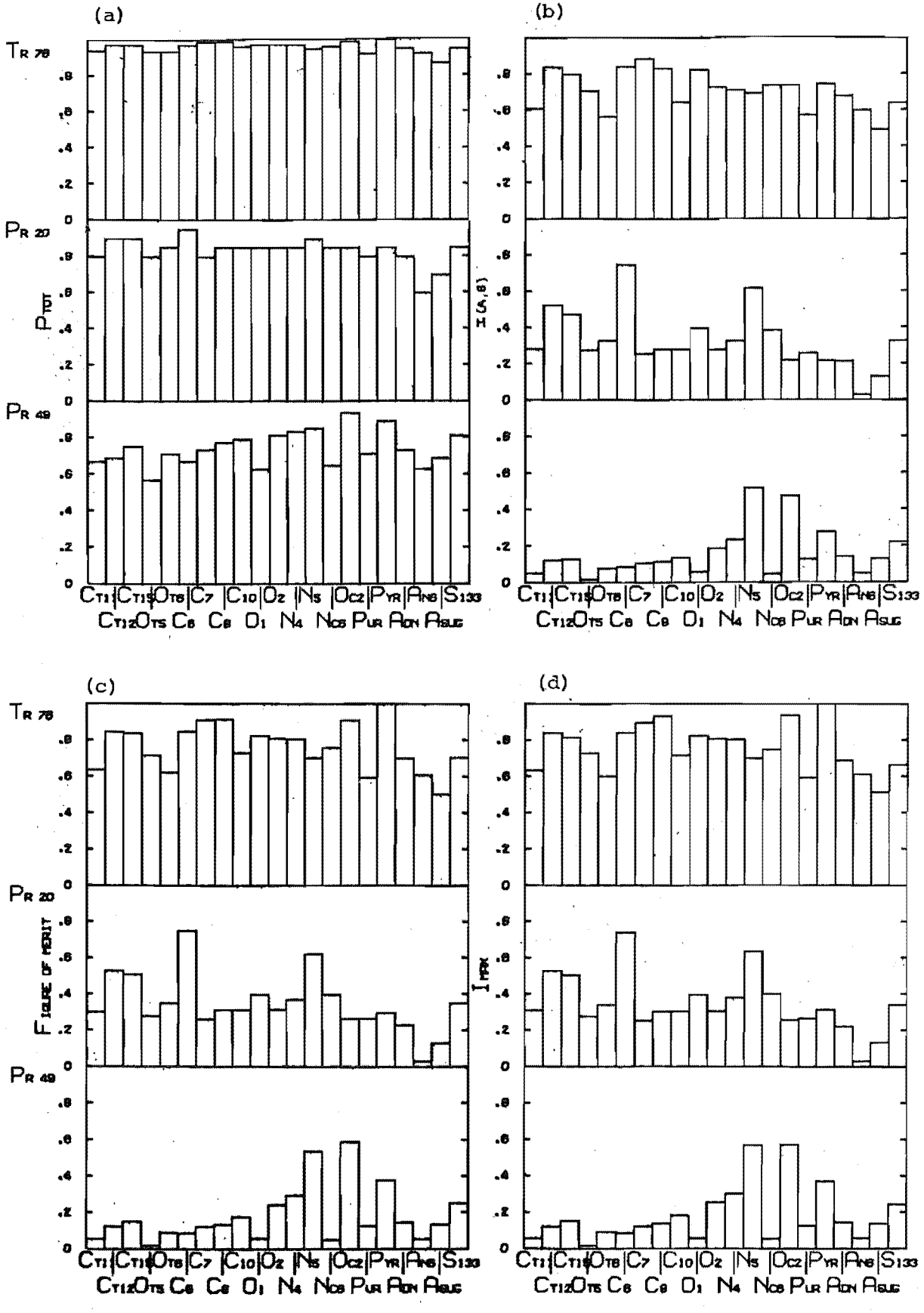
6.4 Discussion

6.4.1 Compositional Effects It is obvious from figure 6.1 that classifier performance decreases from the training to the prediction sets. This is to be expected and the predictions obtained are at least comparable with similar literature values. For example, Wilkins et al. [141] have applied the linear learning machine and sequential simplex optimisation procedures to a much larger (1252) spectral set. Their best values for I(A,B) lay in the range 0.05 - 0.18 with one value of 0.37 i.e. generally less than the average values of 0.324 (table 6.3) and 0.155 (table 6.4) reported here for the two prediction sets. Their best figure of merit values however lay in the range 0.21-0.63 and averaged 0.42, somewhat better than the average of 0.345 reported here for the first prediction set and significantly better than that of 0.180 for the second. This is largely a reflection of the heavily biased compositions of their test sets, the a priori class probabilities lying in all save one case between 0.02 and 0.10. The entropy function (subsection 5.3.2) in the denominator of the expression for figure of merit, a rearranged form of equation (5.3.4)

$$H(A) = -p(1)\log_2 p(1) - p(2)\log_2 p(2) \qquad (6.4.1)$$

returns a value of 0.285 for $p(1) = 0.05$, of 0.880 for $p(1) = 0.30$, and of 1.000 for $p(1) = 0.50$. Consequently similar information gains on two differently composed test sets can lead to widely different M values, the set with the greater disparity between the class sizes giving the higher M. This also explains the similarity between I(A,B) and M for

Figure 6.1: Statistical linear discriminant function analysis using 8-14 m/z values. Histograms of (a) P_{tot}, (b) I(A,B), (c) figure of merit and (d) I_{max} for training and prediction sets.


(a)
(b)
(c)
(d)
TR 78
PR 20
PR 49
PTOT
I(A,B)
FIGURE OF MERIT
IPPK
CT1 CT12 CT15 OT5 OT8 C6 C7 C8 C10 O1 O2 N4 N5 NC8 OC2 PUR PYR AON ANE ASUG S133

the test sets reported here, as it is only for grossly different group sizes (e.g. p(1) = 0.05) that H(A) will differ much from unity.

6.4.2 Individual Analyses It is difficult to identify consistent trends in the results presented. As the relative order of the classifiers within each set is by and large maintained no matter which evaluation measure (P_{tot}, I(A,B), M, or I_{max}) is considered, attention will be focused on M (figure 6.1(c)) as theoretically the most reliable. The second prediction set (Pr49) being larger than the first (Pr20) will perhaps yield the most valid conclusions.

One trend which does emerge from the bottom histogram of figure 6.1(c) is the rise in performance of the elemental composition classifications as the atom number is increased. Thus the classification obtained for total carbon number CT11 is less than that for CT12 which is less than that for CT15 (the first three columns of the histogram). A similar trend is exhibited by OT5 and OT6, and by C6, C7,C8,C10 and O1,O2 and N4, N5 in the base alone. Why ease of classification should increase with increasing atom number is unclear, and it may merely be that for some as yet unexplained reason classification becomes easier as the size of the group to be classified diminishes. Credence is lent to this latter rationalisation by the fact that two other pairs of closely related classification, NC6 and OC2 , and Pur and Pyr, both of which have the number of class members in the first category greater than that in the second, show the same trend. This trend is more weakly shown in Pr20 (the middle histogram of figure 6.1(c)) and does not appear at all in the training set Tr76.

Prediction on carbon number (cf. the bottom histogram of figure 6.1(c)) is essentially invariant whether applied to the total nucleoside (CT11, CT12, CT15) or to the base portion alone (C6, C7, C8, C10). N5 returns the most consistently successful prediction on each of the three sets. The only category for which 100% recognition on the training set was achieved (Pyr) performed well but not outstandingly so on the prediction sets. As others have concluded [147] it is difficult with present knowledge to anticipate performance of a given classification function, or to explain differences in performance among similarly constituted categories.

6.4.3 Performance Factors Of some concern is the significant drop in performance between the two prediction sets, the second being a more extensive selection than the first. Balanced against this however is the fact that the results obtained on either of the prediction sets, when considered independently of one another, indicate quite a reasonable level of classifier performance. This holds whether they are measured by P_{tot}, improvement over more populous category, or by any of the various information gains. The performance difference may perhaps arise from:

(a) lack of true randomness of the training and prediction sets,

(b) too small a training set,

(c) erroneous data base,

(d) gross linear non separability of the classes,

(e) over-training of the classifiers, or

(f) non applicability of the statistical linear discriminant function method.

These factors are dealt with as follows.

The nature of the training set (items a,b and c) is discussed more fully in subsection 5.2.1 and can only really be evaluated by comparing the results obtained here with those obtained on a much larger set of nucleoside spectra. Such a set is not at present available. It was furthermore necessary to exercise a certain amount of selection over the training and the first prediction sets in order to ensure a reasonable class sample in each. Linear separability (item d) is indicated by the results of the simple linear learning machine algorithm (chapter 7) which on 24 features achieved separation for only eleven of the data classes on the training set. With fewer i.e. 8-14 features however, the linear learning machine failed to converge for all but four categories (C7, N4, OC2 and Pyr) and in approximately 12-dimension space the other seventeen classes must be seen as linearly inseparable. The statistical discriminant function used here though depends on the mean of the class rather than more explicitly upon every member as is required for complete linear separability, and consequently this may not be an important factor for this method.

The question of overtraining [155] (item e) is an interesting one even with such a small feature space as that used here. One reason for recording the twenty-one analyses with the full 24 mass positions (tables 6.5-6.7) is to illustrate this phenomenon. These results, as measured

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(1|J)	P(2|N)	I(A,B)	FIG MER	IMAX
1	CT11	47	.618	.957	.931	.947	0.329	.957	.931	.664	.692	.692
2	CT12	35	.461	.971	.999	.987	0.447	.999	.976	.906	.910	.906
3	CT15	29	.382	.966	.999	.987	0.368	.999	.979	.867	.904	.891
4	OT5	42	.553	.881	.971	.921	0.368	.974	.868	.623	.628	.635
5	OT6	25	.329	.880	.961	.934	0.263	.917	.942	.565	.619	.611
6	C6	35	.461	.943	.999	.974	0.434	.999	.953	.842	.846	.840
7	C7	31	.408	.999	.999	.999	0.408	.999	.999	.975	.999	.999
8	C8	25	.329	.999	.999	.999	0.329	.999	.999	.914	.999	.999
9	C10	23	.303	.957	.981	.974	0.276	.957	.981	.712	.805	.803
10	O1	39	.513	.949	.999	.974	0.461	.999	.949	.850	.850	.852
11	O2	24	.316	.958	.999	.987	0.303	.999	.981	.806	.895	.874
12	N4	53	.697	.999	.999	.999	0.303	.999	.999	.884	.999	.999
13	N5	43	.566	.977	.939	.961	0.395	.955	.969	.749	.758	.754
14	NC6	45	.592	.978	.968	.974	0.382	.978	.968	.800	.821	.820
15	OC2	19	.250	.999	.982	.987	0.237	.950	.999	.736	.907	.936
16	PUR	46	.605	.957	.933	.947	0.342	.957	.933	.672	.695	.694
17	PYR	16	.211	.999	.999	.999	0.211	.999	.999	.742	.999	.999
18	ADN	45	.592	.978	.871	.934	0.342	.917	.964	.632	.648	.637
19	AN6	33	.434	.970	.930	.947	0.382	.914	.976	.704	.713	.718
20	ASUG	32	.421	.969	.818	.882	0.303	.795	.973	.519	.529	.541
21	S133	25	.329	.840	.999	.947	0.276	.999	.927	.642	.702	.664
	AV.			.959	.966	.965	0.341	.965	.965	.753	.806	.803

Table 6.5: Statistical discriminant function analysis on Tr76 using 24 m/z values. For column headings see subsection 6.3.2.

Table 6.6: [Overleaf] Statistical discriminant function analysis on Pr20 using 24 m/z values. For column headings see subsection 6.3.2.

Table 6.7: [Overleaf] Statistical discriminant function analysis on Pr49 using 24 m/z values. For column headings see subsection 6.3.2.

Table 6.6

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(2|N) P(1|J)		FIG I(A,B) MER	IMAX
1	CT11	13	.650	.769	.857	.800	0.150	.909	.667	.279	.299	.309
2	CT12	9	.450	.889	.909	.900	0.350	.889	.909	.525	.528	.528
3	CT15	7	.350	.714	.923	.850	0.200	.833	.857	.325	.348	.341
4	OT5	11	.550	.818	.778	.800	0.250	.818	.778	.273	.275	.275
5	OT6	7	.350	.999	.999	.999	0.350	.999	.999	.934	.999	.999
6	C6	9	.450	.778	.636	.700	0.150	.636	.778	.129	.130	.131
7	C7	8	.400	.500	.917	.750	0.150	.800	.733	.163	.168	.164
8	C8	6	.300	.833	.999	.950	0.250	.999	.933	.616	.699	.655
9	C10	6	.300	.667	.929	.850	0.150	.800	.867	.276	.313	.305
10	O1	10	.500	.900	.800	.850	0.350	.818	.889	.397	.397	.397
11	O2	6	.300	.667	.929	.850	0.150	.800	.867	.276	.313	.305
12	N4	14	.700	.857	.833	.850	0.150	.923	.714	.325	.369	.379
13	N5	11	.550	.818	.556	.700	0.150	.692	.714	.112	.113	.112
14	NC6	12	.600	.833	.875	.850	0.250	.909	.778	.385	.397	.402
15	OC2	5	.250	.800	.999	.950	0.200	.999	.938	.541	.667	.610
16	PUR	12	.600	.917	.750	.850	0.250	.846	.857	.361	.372	.367
17	PYR	4	.200	.999	.875	.900	0.100	.667	.999	.446	.618	.717
18	ADN	13	.650	.923	.714	.850	0.200	.857	.833	.325	.348	.341
19	AN6	9	.450	.444	.636	.550	0.000	.500	.583	.005	.005	.005
20	ASUG	9	.450	.999	.636	.800	0.250	.692	.999	.414	.417	.430
21	S133	7	.350	.714	.923	.850	0.200	.833	.857	.325	.348	.341
	AV.			.802	.832	.831	0.202	.820	.836	.354	.387	.386

Table 6.7

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(2|N) P(1|J)		FIG I(A,B) MER	IMAX
1	CT11	34	.694	.706	.667	.694	0.000	.828	.500	.088	.099	.103
2	CT12	24	.490	.583	.800	.694	0.184	.737	.667	.115	.115	.115
3	CT15	14	.286	.714	.829	.796	0.082	.625	.879	.193	.223	.229
4	OT5	28	.571	.536	.619	.571	0.000	.652	.500	.017	.017	.017
5	OT6	14	.286	.714	.829	.796	0.082	.625	.879	.193	.223	.229
6	C6	22	.449	.727	.667	.694	0.143	.640	.750	.114	.115	.116
7	C7	15	.306	.667	.676	.673	-.020	.476	.821	.074	.083	.087
8	C8	13	.265	.692	.889	.837	0.102	.692	.889	.229	.274	.275
9	C10	11	.224	.545	.868	.796	0.020	.545	.868	.109	.143	.146
10	O1	27	.551	.704	.636	.673	0.122	.704	.636	.085	.085	.086
11	O2	11	.224	.909	.868	.878	0.102	.667	.971	.354	.461	.498
12	N4	37	.755	.811	.833	.816	0.061	.938	.588	.244	.303	.325
13	N5	30	.612	.867	.737	.816	0.204	.839	.778	.279	.290	.289
14	NC6	33	.673	.667	.688	.673	0.000	.815	.500	.081	.089	.093
15	OC2	12	.245	.833	.973	.939	0.184	.909	.947	.474	.590	.571
16	PUR	28	.571	.786	.571	.694	0.122	.710	.667	.098	.100	.099
17	PYR	10	.204	.999	.974	.980	0.184	.909	.999	.631	.865	.914
18	ADN	29	.592	.862	.550	.735	0.143	.735	.733	.141	.144	.143
19	AN6	17	.347	.647	.594	.612	-.041	.458	.760	.038	.041	.042
20	ASUG	19	.388	.526	.667	.612	0.000	.500	.690	.026	.027	.028
21	S133	15	.306	.467	.794	.694	0.000	.500	.771	.049	.055	.056
	AV.			.713	.749	.746	0.080	.691	.752	.173	.207	.212

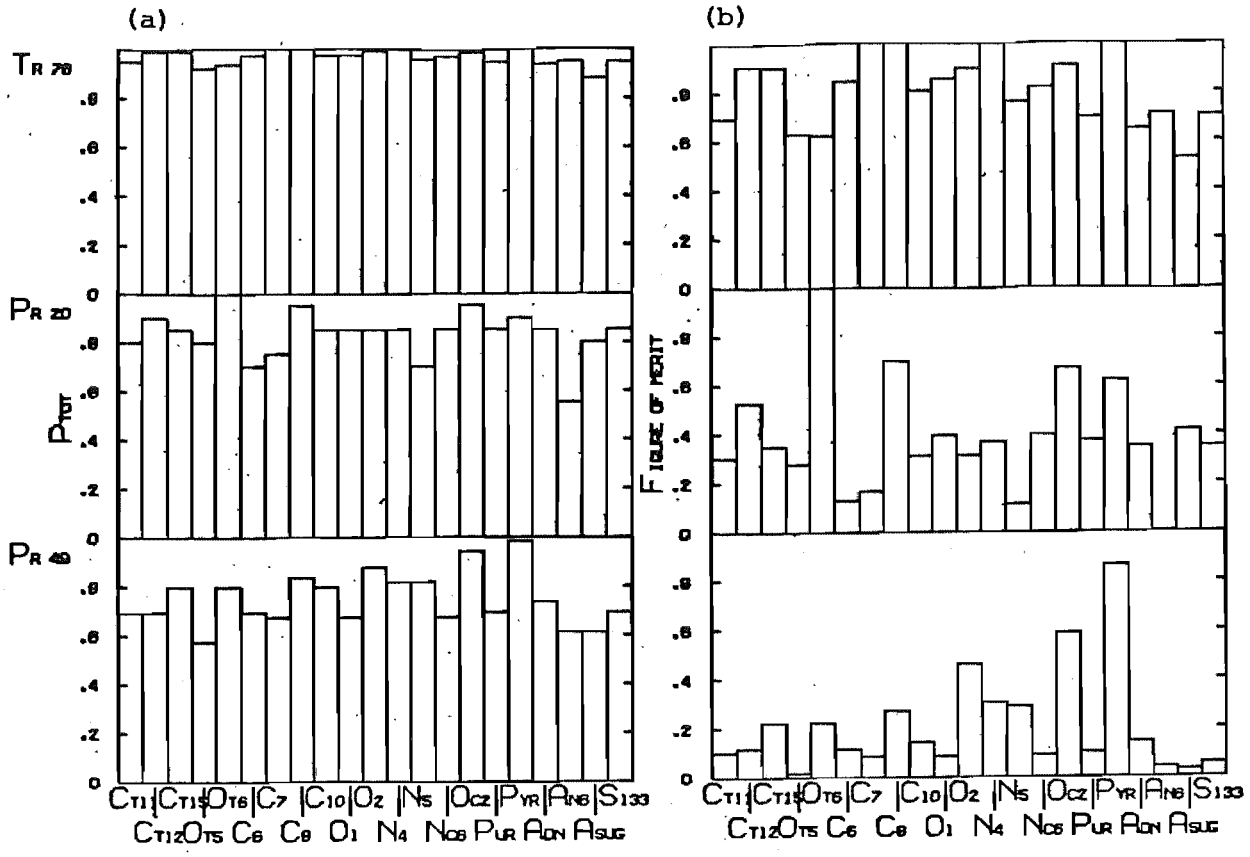


Figure 6.2: Statistical linear discriminant function analysis using 24 m/z values. Histograms of (a) P_{tot} and (b) figure of merit for training and prediction sets.

by P_{tot} and the figure of merit, are depicted in figure 6.2 for the same three data sets. As can be seen by comparison of, say, the bottom histograms of figures 6.1(c) and 6.2(b), prediction is only marginally improved over the reduced feature space classification. This can also be seen by comparison of the average figure of merit values of tables 6.4 (0.180) and 6.7 (0.207).

The method (item f) described in section 6.2, given the grossly non-Gaussian mass spectral intensity distribution, performs equally as well as the error correction feedback technique (chapter 7) and considerably better than the k-nearest neighbour approach (chapter 9) on this data base. A full comparison with these other methods is delayed until chapter 10, but the statistical linear discriminant function method is certainly shown to be statistically very robust.

Mitigating against the fall off in prediction from Pr20 to Pr49 is the consideration that if only one of the prediction sets, even Pr49, had been reported, the results would have been at least acceptable even if not indicative of completely accurate classification. One wonders how many of the classification schemes reported in the literature, often on restricted data bases such as this, would behave if applied to a second prediction set drawn from fresh sources.

6.4.4 Weight Vector Composition Finally, a comment on the weight vectors for each of the twenty-one classes should be made. These are summarised in table 6.8 and reported in full in appendix III. It is generally difficult to comprehend the significance of any but a prominent few of the mass positions comprising the weight vector, especially for elemental compositions, and such comprehension has seldom been attempted in similar studies. The following points however are noteworthy. For sugar = 133 (D- ribose, category 21) m/z 133 is used, and 117 and 146, both with negative coefficients, serve to distinguish the two other most common sugar types deoxyribose (mass = 117) and methylribose (mass = 147, with a characteristic ion at 146 cf. subsection 3.2.1). Unsubstituted adenine has formula $C_5H_5N_5$ and the analyses for base carbon number $C \geqslant 6,7,8,10$ might be expected to reflect substituted adenosine mass values. Thus 119, 120 and 134 appear in C6, 108 in C7, and 120 in C8 and C10 (cf. subsection 3.2.1). Both the base oxygen analyses O1 and

	CT11	CT12	CT15	OT5	OT6	C6	C7	C8	C10	O1	O2
1	111	112	112	112	112	119	108	120	120	112	120
2	112	134	121	115	115	120	126	160	160	115	127
3	117	136	125	121	117	134	160	166	166	126	133
4	120	160	160	126	126	160	163	220	190	134	139
5	133	163	190	135	133	163	190	225	220	151	141
6	165	165	218	151	135	169	191	232	226	164	164
7	169	183	225	169	139	218	218	248	232	169	169
8	170	185	248	219	154	248	225	316	248	219	211
9	194	190	266	249	169	280	248		280	221	218
10	228	218	316	316	220	316	316		316	316	228
11		225			228						232
12		248			316						316
13		280									
14		316									

(Cont.)	N4	N5	NC6	OC2	Pur	Pyr	Adn	AN6	Asug	S133
1	112	112	115	112	112	112	112	112	112	109
2	115	115	117	125	125	115	119	120	127	117
3	125	126	134	127	127	125	125	126	136	133
4	126	136	135	140	141	127	127	135	149	146
5	127	148	136	151	148	146	134	160	162	149
6	135	164	139	168	177	185	151	169	169	160
7	152	179	141	169	191	208	169	171	202	162
8	169	183	146	228	218	211	170	194	218	164
9	171	202	155				179	211		202
10	184	208	164				185	218		219
11	218	248	166				218			
12	228		178				232			
13			179							
14			248							

Table 6.8: Most discriminatory mass positions. The 8-14 m/z values used in the statistical discriminant function analyses of various categories of nucleoside mass spectra.

O2 display negative coefficients for their only mass value characteristic of oxygen deficient adenine m/z 164 (Ade +30). The analysis for N at C6 shows the characteristic adenosine m/z values 134, 135, 136, 164 and 178 (Ade +44) and that for O at C2 gives 112 (Cyt + 2 or Ura + 1). The other analyses fail to show any significant mass positions which could be explained by existing mass spectral knowledge.

In conclusion, then, the results obtained using the statistical linear discriminant function method are at least comparable to those of similar studies reported in the literature using other methods such as k-nearest neighbour and linear learning machine. The classifiers developed gave reasonable categorisation on unknown spectra. A definitive evaluation of the method however must await a more extensive data base and this is not at present available.

Chapter 7

LEARNING MACHINE APPROACH

7.1 Introduction

A linear learning machine or error correction feedback algorithm was applied to the data base described in subsection 5.2.1. The method entails complete linear separation of class members and non members in the hyperspace chosen. This is done by repeated modification of the weight vector by any misclassified spectra until the entire training set is classified correctly, and is described in more detail in section 7.2. The method has enjoyed wide applicability in chemical [281] and especially in mass spectrometric [282,141] problems and a comparison with similar literature studies is made (subsection 7.3.3).

7.2 Method

A simple binary linear learning machine program employing error correction feedback has been compiled by Jurs and Isenhour [283] and was used together with suitable input routines in this work. A brief description of the heart of the program follows. An initially arbitrary classification function i.e. discriminant surface was trained on the set of 76 spectra (Tr76). Class membership was denoted by +1 for members and by -1 for class non members. Training was accomplished by classification of and modification if necessary by each point (spectrum) in turn in the n-dimensional hyperspace. The particular error correction feedback technique employed involved what could be geometrically described as a reflection of the decision surface about each misclassified point. On the first pass a record is kept of those patterns misclassified, i.e. those which modified the decision surface, and on the second pass only these are classified by the new weight vector. This continues until the set of misclassified spectra vanishes. The weight vector resulting is applied afresh to the full training set and the whole process begun again. This continues until either all the patterns are correctly classified by the one (final) weight vector, or until a limit of 1000 passes is reached in which case the training procedure is deemed not to have converged and the classification attempt is terminated. The number of feedbacks required for

each convergent analysis (table 7.1) lay in the range 17-200, with the exception of CT15 which when used with a deadzone required 666 and 294 feedbacks for the logarithmic and binary data forms respectively (see below). Thus the classifier is repeatedly asked a series of questions

		Log		Bin	
Deadzone:		0.0	0.1	0.0	0.1
1	CT15	98	666	121	294
2	C6	24	36	30	63
3	C7	25	36	56	132
4	C8	26	44	-	-
5	C10	45	109	57	86
6	O1	25	57	40	196
7	O2	22	74	27	80
8	N4	17	20	28	33
9	OC2	19	30	19	46
10	Pyr	18	26	22	46
11	AN6	52	113	-	-

Table 7.1: Number of feedbacks for linear learning machine training on Tr76 using 24 m/z positions. Autoscaled logarithmic and binary data with and without deadzone of 0.1.

until it responds correctly to all of them. It is this improvement in performance with "experience" that has led to the use of the term "learning machine".

Reflection about each misclassified point is expressed in matrix notation by

$$\underline{W}_1 = \underline{W} + c\underline{X}_i \qquad (7.2.1)$$

where $\underline{W}_1$ is the new weight vector, $\underline{W}$ the old, $\underline{X}_i$ the misclassified pattern vector and c the (scalar) correction increment, which in this simple implementation of the process is the same for all mass positions. It is calculated from

$$c = \frac{-2s}{\underline{X}_i' \underline{X}_i} \qquad (7.2.2)$$

where s is the scalar product of the original weight vector $\underline{W}$ with the misclassified pattern

$$s = \underline{W}\,\underline{X}_i \quad . \qquad (7.2.3)$$

The sign of s determines in which class the pattern is placed i.e. for $\underline{X}_i$ to be misclassified s has the wrong sign. The improved weight vector necessarily gives a scalar product s_1 of the correct sign with the previously misclassified pattern

$$s_1 = \underline{W}_1 \underline{X}_i \quad . \qquad (7.2.4)$$

To reduce the dimensionality of the problem from a possible maximum of 755 (m/z 1-755) a feature selection process was employed. This was simply to select mass positions according to their F ratios as computed by the statistical linear discriminant algorithm described in chapter 6. Thus those mass positions which reflected the greatest differences between the two groups (class members and non members) and the greatest similarities within each group were selected. An arbitrary number of 24 mass positions was selected after initially investigating the use of only 8-14, which latter gave convergence in only four of the twenty-one categories.

Two forms of spectra as described in subsection 5.2.2 were investigated, autoscaled logarithmic spectra, i.e. the same data base as used for the statistical linear discriminant analyses of chapter 6, and binary (peak/no peak) spectra with the same intensity threshold of 1% of the base peak. It could be expected that spectra in the hyperspace lying very close to the decision surface, i.e. having scalar products close to zero, would be difficult to classify. To examine such cases a deadzone or region of non classification of $\pm$ 0.1 was applied to both the autoscaled logarithmic and the binary spectra analyses.

Total CPU time for the analysis of each category, i.e. time for convergence of the weight vector and for prediction on Pr49, was 10-15 sec with 24 m/z values and 6-10 sec with 8-14 m/z values.

7.3 Results and Discussion

7.3.1 Presentation of Results In ten of the twenty-one categories convergence on the training set was not achieved i.e. they were shown to be linearly inseparable in the chosen 24-dimension hyperspaces by this method. When the data base was reduced to binary form a further two categories, C8 and AN6, also failed to converge. Results for the four kinds of analysis, logarithmic and binary data with and without a deadzone, are summarised in table 7.2 by the average P_{tot} and figure of merit M values. These are averages over the eleven convergent categories for the logarithmic case, and over nine for the binary.

As discussed more fully in subsection 7.3.2 logarithmic data

		Log		Bin	
		Pr20	Pr49	Pr20	Pr49
		P_{tot} (av.)			
Ddz	0.0	84.5%	75.0%	87.8%	76.4%
	0.1	87.0	78.7	89.2	74.5
		M (av.)			
Ddz	0.0	0.469	0.237	0.490	0.253
	0.1	0.556	0.281	0.588	0.223

Table 7.2: Average P_{tot} and figure of merit values for the four linear learning machine methods. 24 m/z values used for each. See subsection 7.3.1.

without a deadzone gave the best overall classification, and for this analysis type the full results for the training and the two prediction sets (Tr76, Pr20, and Pr49) are reproduced in tables 7.3 - 7.5 respectively. These are the results taken as "best" for the comparisons of chapter 10. This analysis type is also presented in the form of histograms for the three test sets in figures 7.1(a) (P_{tot}) and 7.2(a) (figure of merit). Results for the other three analysis types

are not reproduced in full but are summarised by the other histograms of figures 7.1 and 7.2. The parts of these are (b) binary data with zero deadzone, (c) logarithmic data with deadzone = 0.1, and (d) binary data with deadzone = 0.1. Note that for those analysis types involving a deadzone the spectra not classified have been omitted from the performance calculations i.e. only those spectra actually assigned one way or the other have been included in P_{tot}, M, etc. This only applies to the prediction sets as for the training set perfect recognition must be achieved (section 7.2). The percentages of spectra not classified after the imposition of this deadzone are listed separately in table 7.6. Full 24-dimension weight vectors are reproduced in appendix III.

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(1|J)	P(2|N)	I(A,B)	FIG MER	IMAX
1	CT15	29	.382	.999	.999	.999	0.382	.999	.999	.959	.999	.999
2	C6	35	.461	.999	.999	.999	0.461	.999	.999	.995	.999	.999
3	C7	31	.408	.999	.999	.999	0.408	.999	.999	.975	.999	.999
4	C8	25	.329	.999	.999	.999	0.329	.999	.999	.914	.999	.999
5	C10	23	.303	.999	.999	.999	0.303	.999	.999	.884	.999	.999
6	O1	39	.513	.999	.999	.999	0.487	.999	.999	.999	.999	.999
7	O2	24	.316	.999	.999	.999	0.316	.999	.999	.900	.999	.999
8	N4	53	.697	.999	.999	.999	0.303	.999	.999	.884	.999	.999
9	OC2	19	.250	.999	.999	.999	0.250	.999	.999	.811	.999	.999
10	PYR	16	.211	.999	.999	.999	0.211	.999	.999	.742	.999	.999
11	AN6	33	.434	.999	.999	.999	0.434	.999	.999	.987	.999	.999
	AV.			.999	.999	.999	0.353	.999	.999	.914	.999	.999

Table 7.3: Linear learning machine analysis on Tr76. 24 m/z values used with autoscaled logarithmic data and zero deadzone For column headings see subsection 6.3.2.

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(2|N) P(1|J)		I(A,B)	FIG MER	IMAX
1	CT15	7	.350	.999	.692	.800	0.150	.636	.999	.414	.443	.485
2	C6	9	.450	.999	.636	.800	0.250	.692	.999	.414	.417	.430
3	C7	8	.400	.875	.917	.900	0.300	.875	.917	.505	.520	.520
4	C8	6	.300	.833	.999	.950	0.250	.999	.933	.616	.699	.655
5	C10	6	.300	.500	.786	.700	0.000	.500	.786	.057	.064	.065
6	O1	10	.500	.800	.600	.700	0.200	.667	.750	.125	.125	.125
7	O2	6	.300	.999	.857	.900	0.200	.750	.999	.557	.632	.689
8	N4	14	.700	.999	.833	.950	0.250	.933	.999	.616	.699	.655
9	OC2	5	.250	.999	.933	.950	0.200	.833	.999	.616	.760	.820
10	PYR	4	.200	.999	.875	.900	0.100	.667	.999	.446	.618	.717
11	AN6	9	.450	.667	.818	.750	0.200	.750	.750	.181	.183	.182
	AV.			.880	.813	.845	0.191	.755	.921	.413	.469	.486

Table 7.4: Linear learning machine analysis on Pr20. 24 m/z values used with autoscaled logarithmic data and zero deadzone. For column headings see subsection 6.3.2.

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(2|N) P(1|J)		I(A,B)	FIG MER	IMAX
1	CT15	14	.286	.857	.629	.694	-.020	.480	.917	.151	.175	.190
2	C6	22	.449	.818	.593	.694	0.143	.621	.800	.131	.132	.133
3	C7	15	.306	.800	.647	.694	0.000	.500	.880	.129	.145	.154
4	C8	13	.265	.692	.806	.776	0.041	.563	.879	.153	.183	.190
5	C10	11	.224	.818	.816	.816	0.041	.563	.939	.223	.291	.313
6	O1	27	.551	.704	.455	.592	0.041	.613	.556	.019	.019	.019
7	O2	11	.224	.909	.684	.735	-.041	.455	.963	.196	.255	.294
8	N4	37	.755	.892	.999	.918	0.163	.999	.750	.538	.670	.744
9	OC2	12	.245	.833	.811	.816	0.061	.588	.938	.244	.303	.325
10	PYR	10	.204	.800	.872	.857	0.061	.615	.944	.248	.339	.359
11	AN6	17	.347	.765	.594	.653	0.000	.500	.826	.088	.094	.098
	AV.			.808	.719	.750	0.045	.591	.854	.193	.237	.256

Table 7.5: Linear learning machine analysis on Pr49. 24 m/z values used with autoscaled logarithmic data and zero deadzone. For column headings see subsection 6.3.2.

		Log		Bin	
		Pr20	Pr49	Pr20	Pr49
1	CT15	0.0%	18.4%	5.0%	16.3%
2	C6	10.0	8.2	25.0	24.5
3	C7	20.0	10.2	5.0	6.1
4	C8	10.0	12.2	-	
5	C10	10.0	6.1	5.0	12.2
6	O1	25.0	22.5	5.0	10.2
7	O2	20.0	16.3	0.0	10.2
8	N4	20.0	10.2	5.0	4.1
9	OC2	10.0	12.2	15.0	16.3
10	Pyr	10.0	12.2	20.0	16.3
11	AN6	15.0	12.2	-	
	Av.	13.6	12.8	9.5	12.9

Table 7.6: Percentages of spectra in prediction sets not classified by linear learning machine analyses with deadzone.

7.3.2 Effect of Pre-processing Almost half the analyses attempted failed to converge - CT11, CT12, OT5, OT6, N5, NC6, Pur, Adn, Asug, S133 and C8 and AN6 in the binary case. It is difficult to postulate a theoretical explanation for this linear inseparability. It may however be noted that the elemental composition classes for the nucleoside as a whole (CT11, CT12, OT5, and OT6) were often shown to be inseparable, as also were the various adenosine base types - Asug, Adn, Pur and NC6 and AN6 with binary data. It is possible that some of these classes may be linearly separable in other 24-dimension spaces which could be found with more sophisticated methods of feature selection.

Reduction to binary spectra has marginal disadvantages. A very slight gain in average performance of 3.3% on the first prediction set (Pr20) and of 1.4% on the second (Pr49), or of +0.021 and +0.016 in terms of the figure of merit (values obtained by subtractions from table 7.2) is offset by the non convergence of two further categories, C8 and AN6. Any computational saving is negligible with such a small data base, although if a larger set of spectra were available it may

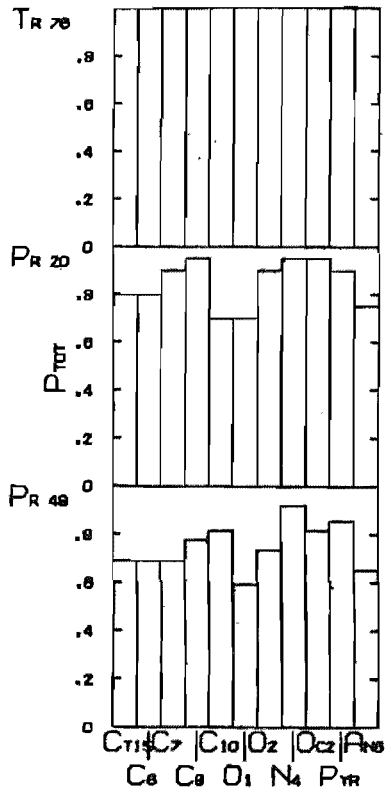
become important. The relative order between the binary and the logarithmic analyses is approximately maintained, save that (compare figure 7.2 (a) and (b)) O2 on Pr20 gives perfect classification with binary data, and CT15 on Pr49 is very much improved (figure of merit 0.175 with logarithmic data (table 7.5 or figure 7.2(a)) and 0.583 (figure 7.2(b)) with binary) over the logarithmic case.

Imposition of a deadzone in the logarithmic case again marginally improves performance, raising the average P_{tot} by 2.5% on Pr20 and by 3.7% on Pr49, or in terms of the figure of merit by +0.087 and +0.044 (table 7.2). The penalty is a considerable percentage of spectra not classified, averaging about 13% (table 7.6), and the very slight increase in prediction ability cannot justify this. On both prediction sets the relative order is again maintained (compare figure 7.2(a) and (c)) with the exception of category C10 on Pr20 which increases the figure of merit value from 0.064 to unity, i.e. perfect classification, when the deadzone is imposed at the expense of 10.0% non classification (table 7.6). This must however be viewed at present as merely a quirk of the data. In the binary analysis the deadzone increases average performance on Pr20 by 1.4% but decreases it on Pr49 by 1.9%, or in terms of the figure of merit by +0.098 and by -0.030 (table 7.2). Again there is 9-13% non classification (table 7.6). The relative order, with minor aberrations, is kept approximately constant (compare figure 7.2(b) and (d)).

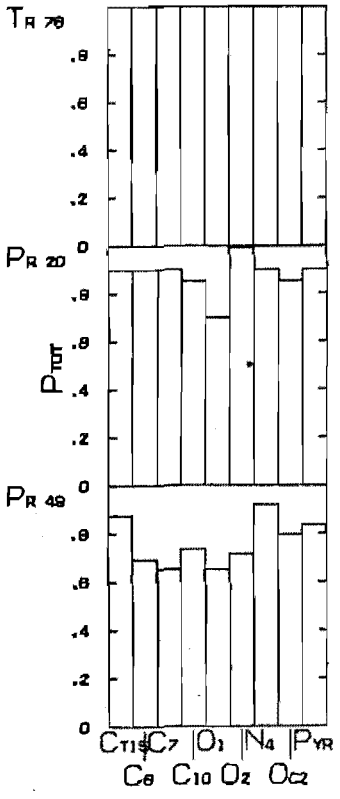
All of these variations in performance induced either by reduction to binary form or by imposition of a region of non classification are of such a minor and/or random nature that very little significance can be attached to them. The autoscaled logarithmic data without a deadzone gives the best convergence and at least equal performance of the four methods. This is in accord with earlier results of Kowalski and Bender [223]. Consequently, and for the sake of uniformity, these results will be taken as typical of the learning machine method on this data set.

7.3.3 Individual Analyses As noted in subsection 6.4.2 for the statistical linear discriminant function analyses it is difficult to identify trends in the results presented. This is accentuated by nonconvergence of many of the categories. One previously identified (cf. subsection 6.4.2) trend of increasing classification success with

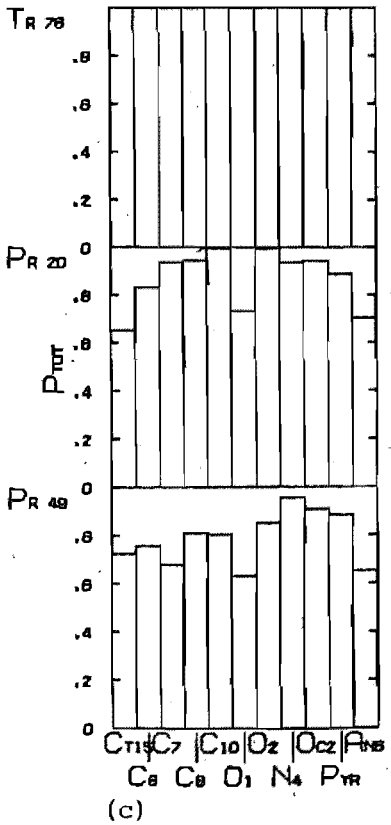
Figure 7.1: Histograms of P_{tot} for linear learning machine analyses. 24 m/z values used with (a) logarithmic data and zero deadzone, (b) binary data and zero deadzone, (c) logarithmic data and deadzone = 0.1, and (d) binary data and deadzone = 0.1.


TR 76
PR 20
PTOT
PR 49
.8
.6
.4
.2
0
CT1$C7 |C10|O2 |OC2|PYR
C8 C9 O1 N4 PYR

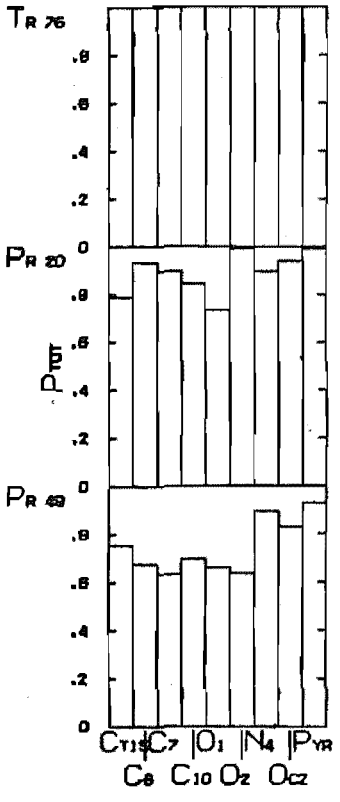
(a)


TR 76
PR 20
PTOT
PR 49
.8
.6
.4
.2
0
CT1$C7 |O1 |N4 |PYR
C8 C10 O2 OC2

(b)

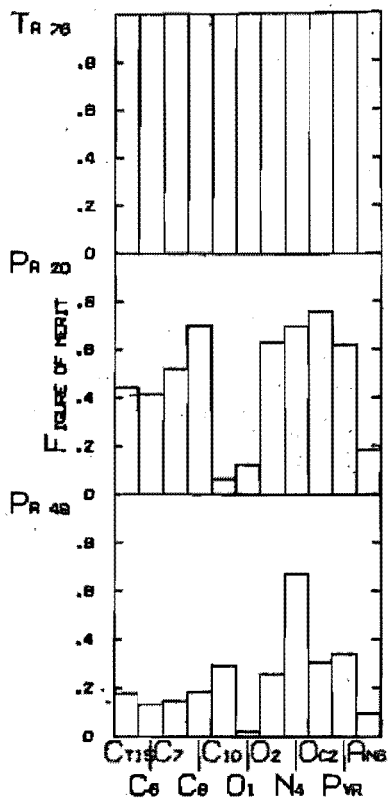

TR 76
PR 20
PTOT
PR 49
.8
.6
.4
.2
0
CT1$C7 |C10|O2 |OC2|PYR
C8 C9 O1 N4 PYR

(c)

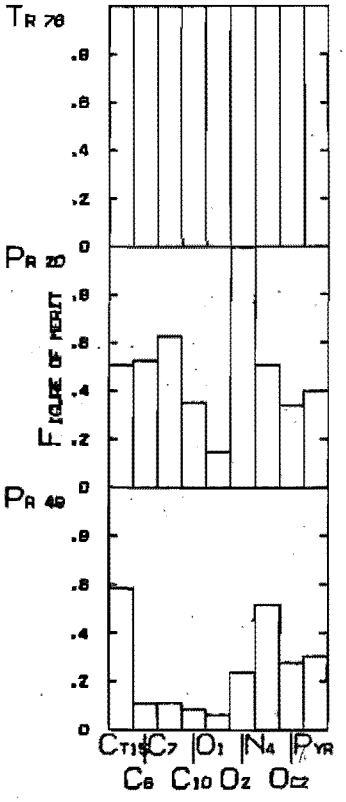

TR 76
PR 20
PTOT
PR 49
.8
.6
.4
.2
0
CT1$C7 |O1 |N4 |PYR
C8 C10 O2 OC2

(d)

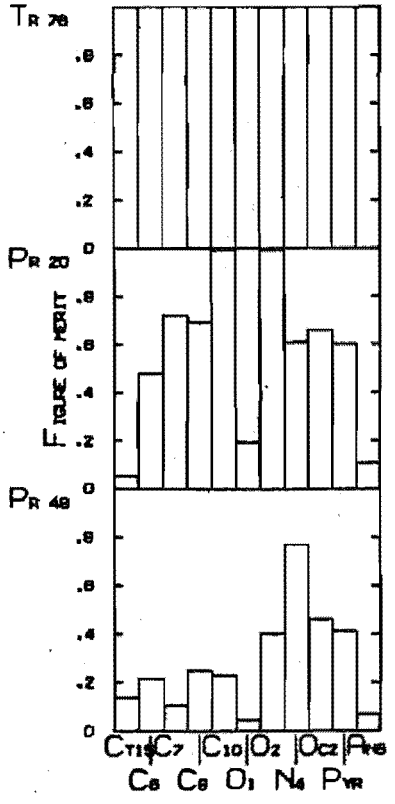
Figure 7.2: Histograms of figure of merit for linear learning machine analyses. 24 m/z values used with (a) logarithmic data and zero deadzone, (b) binary data and zero deadzone, (c) logarithmic data and deadzone = 0.1, and (d) binary data and deadzone = 0.1.


TR 78
PR 20
PR 48
FIGURE OF MERIT
CT1§C7 |C10|O2 |OC2|PNB
C8 C9 O1 N4 PYR

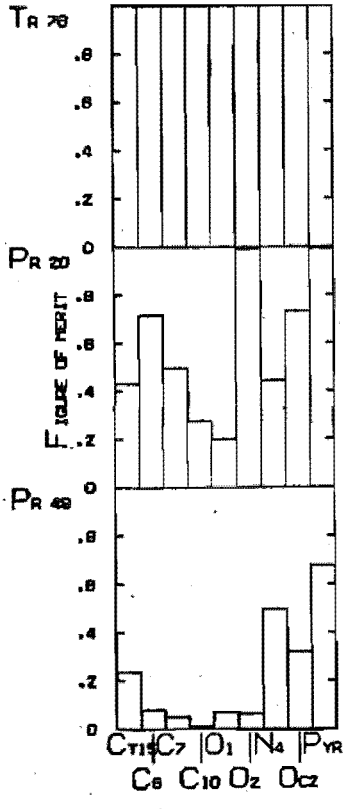
(a)


TR 78
PR 20
PR 48
FIGURE OF MERIT
CT1§C7 |O1 |N4 |PYR
C8 C10 O2 OC2

(b)


TR 78
PR 20
PR 48
FIGURE OF MERIT
CT1§C7 |C10|O2 |OC2|PNB
C8 C9 O1 N4 PYR

(c)


TR 78
PR 20
PR 48
FIGURE OF MERIT
CT1§C7 |O1 |N4 |PYR
C8 C10 O2 OC2

(d)

increasing atom number in the compositional analyses is again evident, as can be seen from figure 7.2(a). Both on Pr20 and on Pr49 the analyses for base carbon number C6, C7, C8 and C10 and for base oxygen number O1 and O2 rise in performance in the order listed. This is however not at all reproduced in the binary or deadzone analyses (figure 7.2(b)-(d)) and its significance and generality must remain dubious. As further noted in subsection 6.4.2 it is difficult and seldom attempted to predict or explain differences in performance even on similarly constituted categories for any pattern recognition method.

The results obtained here compare well with similar studies by other workers. The same linear learning machine method and a sequential simplex optimisation procedure used by Wilkins et al. [141] on a much larger data base gave for several analyses an average figure of merit value of 0.42. Rotter and Varmuza in their steroid study [145] using a distance from the mean classification technique did not adopt the figure of merit, although this measure can be calculated from their results and yields an average of 0.35. Both these studies compare very favourably with the average figure of merit value obtained here of 0.469 on the first prediction set, and somewhat less favourably with that of 0.237 on the second (table 7.2). Reasons for the drop in performance between the two prediction sets are discussed in subsection 6.4.3. Aside from the obvious linear inseparability of some of the categories, the size of the test set makes a definitive evaluation difficult. However, even if only the result on Pr49 had been reported the classifications achieved are at least comparable with similar work reported in the literature.

Chapter 8

DISTANCE FROM THE MEAN CLASSIFICATION

8.1 Introduction

One of the simplest binary classification techniques is to represent each class by an average vector, and to assign an unknown according to its distance from these two means. This method possesses the advantages of being conceptually trivial, applicable to any number of mass positions, and computationally extremely rapid once the means have been established. Consequently it has been extensively applied to other mass spectral problems [143,125b,144] including the steroid work of Rotter and Varmuza [145]. In this present work the method has been applied to the data base described in subsection 5.2.1 and presented to the classifiers in the forms described in subsection 5.2.2. The results obtained (section 8.3) are comparable with those of similar studies by other workers [145].

On the present data base this approach performs the best of the four pattern recognition methods, but a detailed comparison is delayed until chapter 10.

8.2 Method

The distance from the mean approach is a standard, conceptually very simple technique which has been adequately described elsewhere [143]. Two mean vectors are found for the patterns, i.e. spectra, in the d-dimensional hyperspace chosen. These are for class members ($\underline{\bar{X}}_1$) and non members ($\underline{\bar{X}}_2$)

$$\underline{\bar{X}}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \underline{X}_{1i} \qquad (8.2.1)$$

$$\underline{\bar{X}}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \underline{X}_{2i} \qquad (8.2.2)$$

where each $\underline{X}_{1i}$ is one of the n_1 patterns in class 1 (class members) and each $\underline{X}_{2i}$ is one of the n_2 patterns in class 2 (class non members). Each

unknown spectrum j of the prediction sets is assigned to one of the two classes according to its generalised euclidean distance

$$D_{kj} = \sqrt{\sum_{\ell=1}^{d} (x_{j\ell} - \bar{x}_{k\ell})^2} \qquad k = 1,2 \qquad (8.2.3)$$

from the two means. $x_{j\ell}$ and $\bar{x}_{k\ell}$ are the ℓth components of the d-dimensional pattern vectors $\underline{X}_j$ and $\underline{\bar{X}}_k$.

To ensure compatibility with the other classification techniques used in this work (chapters 6,7 and 9) and independence of the training and prediction processes, the means were calculated on the training set

		82 m/z		24 m/z		8-14 m/z		
		log	bin	log	bin	orig	log	bin
1	CT11	63	59	23	18	10	10	9
2	CT12	68	58	24	19	14	14	12
3	CT15	63	58	22	20	10	8	8
4	OT5	62	60	23	18	10	10	8
5	OT6	63	61	22	17	12	12	9
6	C6	67	60	22	21	10	9	10
7	C7	68	59	23	20	10	9	9
8	C8	67	60	22	20	8	8	6
9	C10	65	57	21‡	20‡	10	9	8
10	O1	66	61	23‡	21‡	10	10	9
11	O2	63	37	20	11	12	10	5
12	N4	53	55	24	21	12	12	10
13	N5	51	43	23	19	10	10	5
14	NC6	68	61	22	17	14	14	10
15	OC2	69	67	24	20	8	8	7
16	Pur	45	43	21	18	8	7	8
17	Pyr	72	66	23	15	8	8	5
18	Adn	47	41	19	17	12	11	7
19	AN6	64	57	19	20	10	9	10
20	Asug	43	58	21	19	8	8	8
21	S133	59	57	21	19	10	8	8
Av.		61.2	56.1	22.0	18.6	10.3	9.7	8.1

‡ 25 rather than 24 m/z positions originally used for category O1.

Table 8.1: Numbers of mass positions used for distance from mean classifications. Approximately equal components have been removed. For the most reduced set (8-14 m/z) starting numbers of components are also tabulated.

Tr76 alone. The spectra of the prediction sets were then classified according to these mean vectors, despite the use by other workers [145] of the one set of spectra not only for calculation of the means but also for use as unknowns to be predicted. In an effort to isolate those vectors lying almost equidistant from the two means, a deadzone of variable magnitude was imposed within which classification of patterns was not attempted. This was adjusted in each case to give 10-15% non classification on the training set, as can be seen from figure 8.2, although the binary results (see below) sometimes exceed this range.

The analyses were conducted with d = 82,24 and 8-14 dimensions i.e. mass positions, using the m/z selections obtained as described in section 6.2. For efficiency of computation those components of the means which were equal, within narrow limits, for the two classes were omitted from the calculations. This reduction in the number of components for each of the three sets of m/z values is reported in table 8.1. There actual numbers of components used for each of the twenty-one categories are tabulated, together with an average dimensionality for each starting set (82, 24 or 8-14 m/z). The dimensionality reduction was in general greater for binary than for logarithmic spectra due to the lower information content of the former. It should furthermore be borne in mind that while there was much overlap between the sets of m/z positions for the various categories, they were in general different sets.

The spectra were presented to the classifiers in the forms described in subsection 5.2.1 and used throughout this work i.e. autoscaled logarithmic and binary spectra with prior normalisation to 100% of the base peak and with retention of only peaks $\geqslant$ 1% relative intensity and $\geqslant$ 100 amu. Both the distance from the mean method and the k-nearest neighbour approach of the following chapter were encoded in program KNNCLASSIF which is reproduced in appendix II. CPU time for each analysis of each category was in all cases under 1 sec.

8.3 Results and Discussion

8.3.1 Variants of Method The twelve types of analysis conducted using the distance from the mean technique are summarised in table 8.2 by average figure of merit and P_{tot} values for

the twenty-one classes described in subsection 5.2.3. The table comprises the results of the classifications using binary and autoscaled logarithmic data with and without a deadzone, for the three reduced m/z selections described in section 8.2. The average numbers of m/z positions from table 8.1 and the non classification effects of the deadzone are also recorded. Only the results for the augmented prediction set, Pr49 are given as most of the discussion will centre on these. The data of table 8.2 is graphed in figure 8.1 where the average figure of merit is plotted as a function of average number of m/z positions used in the analyses i.e. as a function of average dimensionality. Four plots are drawn for the logarithmic and binary data with and without a deadzone.

As can be seen from figure 8.1 the logarithmic spectra with a deadzone and using an average of 61.2 m/z positions gave the best overall results (average M = 0.391). However this was achieved at the expense of an average of 24% of spectra not classified (the fourth column of table 8.2) and such a high percentage of unassigned spectra must render such a classification method of little use. Consequently the same logarithmic spectra again using an average of 61.2 m/z values, but without

(Pr49)		No ddz.		Ddz.		
	Av. Dim.	M	P_{tot}	Not Classif.	M	P_{tot}
log $\leqslant$ 82	61.2	.273	79.0%	24.4%	.391	83.0%
$\leqslant$ 24	22.0	.192	75.4	17.1	.285	80.6
$\leqslant$ 8-14	9.7	.142	72.9	7.9	.169	74.9
bin $\leqslant$ 82	56.1	.160	71.0%	48.9%	.268	82.4%
$\leqslant$ 24	18.6	.145	70.5	32.2	.214	75.8
$\leqslant$ 8-14	8.1	.132	67.3	15.9	.159	67.7

Table 8.2: Distance from mean classification success on Pr49. Results for binary and autoscaled logarithmic data with three sets of m/z values (section 8.2). Columns show the average dimensionality, the figure of merit M and P_{tot} averages for classification with and without a deadzone, and the average percentage not assigned for the deadzone case.

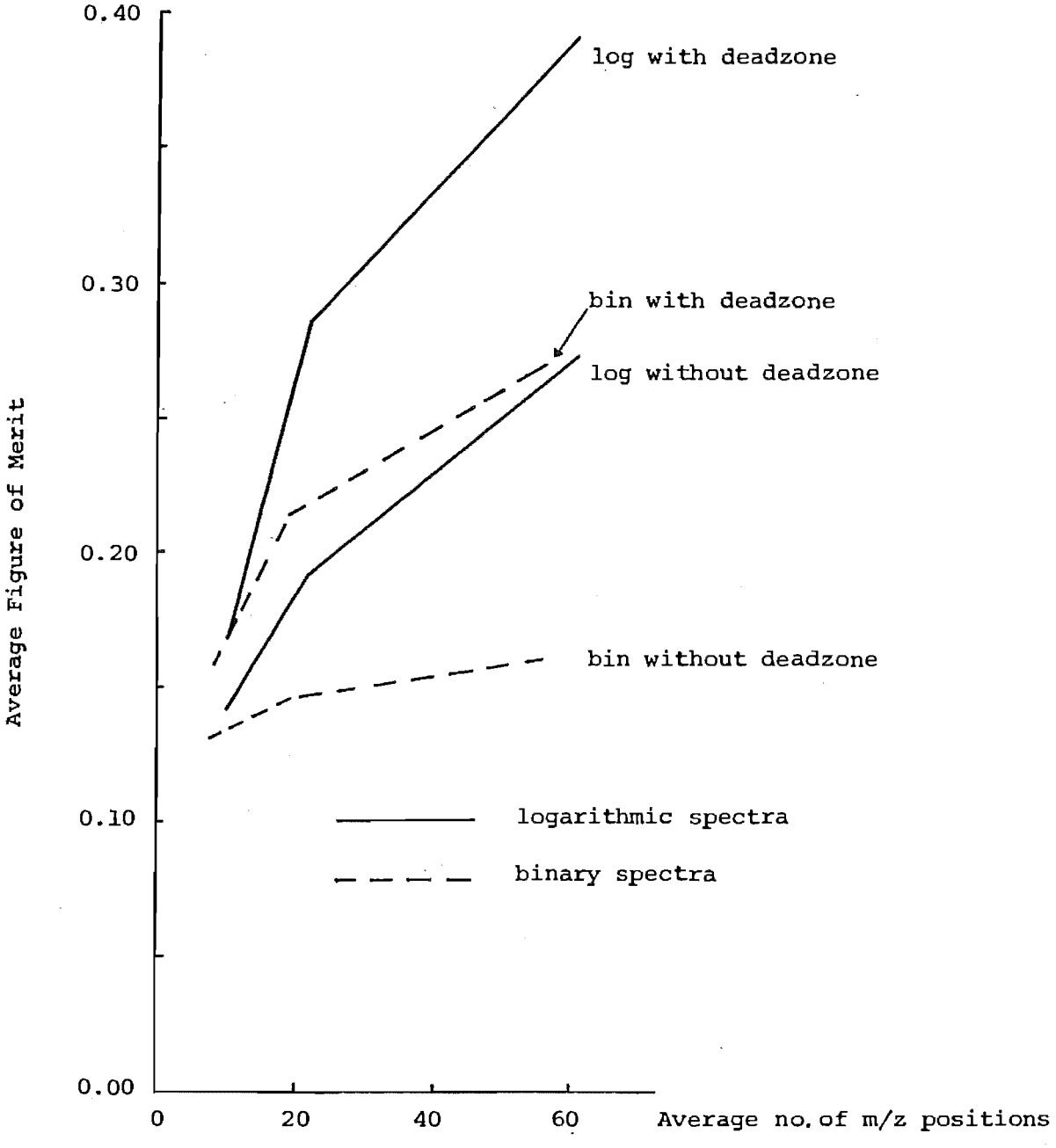


Figure 8.1: Graph of average distance from mean classification success. Average figure of merit plotted against average number of m/z positions used for binary and logarithmic data with and without a deadzone, on the second prediction set Pr49.

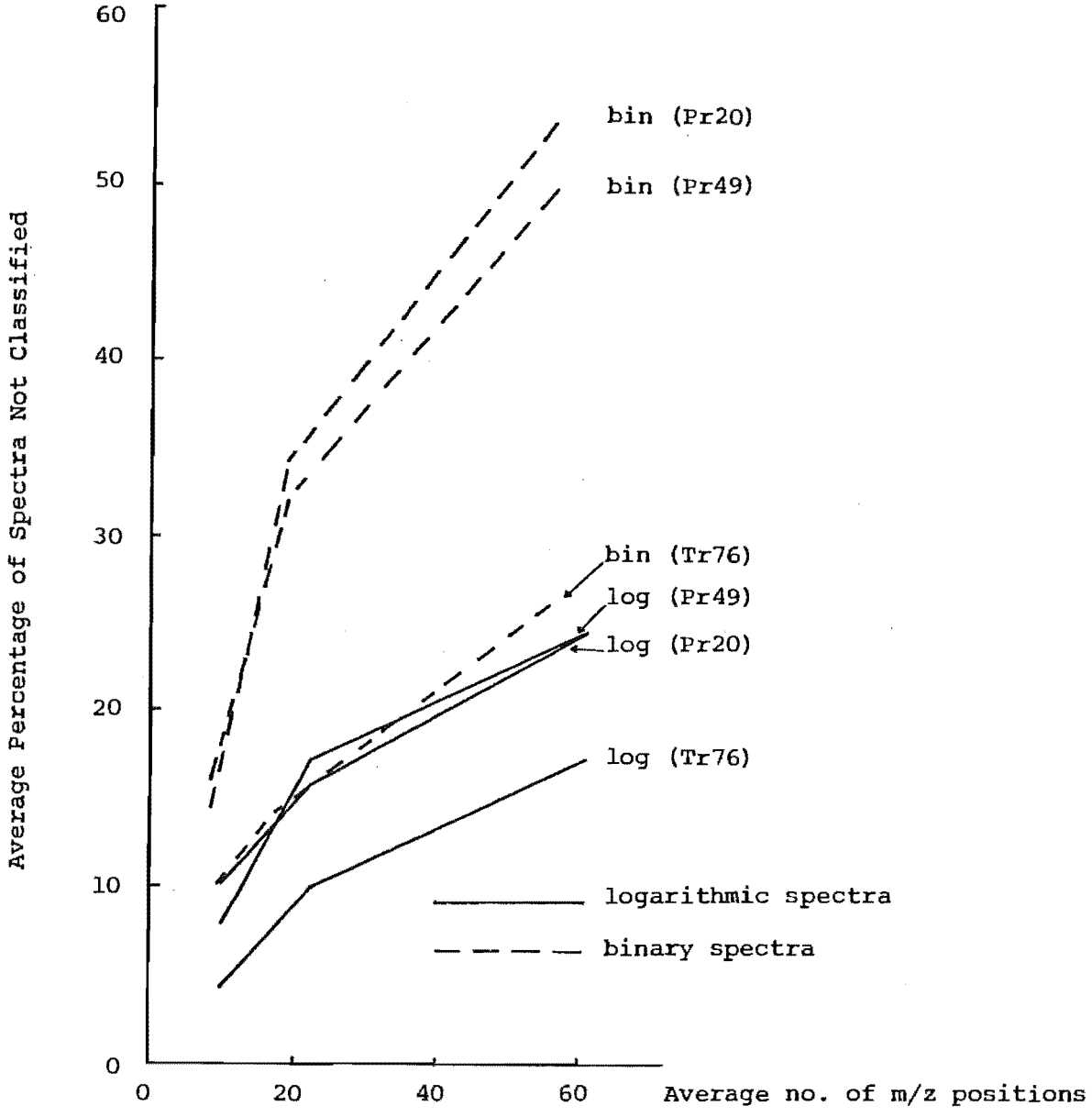


Figure 8.2: Average percentages of spectra not assigned by distance from mean classifications with deadzone. Binary and logarithmic data forms plotted as a function of average number of m/z positions used, for the training and the two prediction sets.

application of a deadzone, has been selected as the "best" variant of the twelve reported here. This form returns an average figure of merit of 0.273 corresponding to an overall success rate of P_{tot} = 79.0% (table 8.2). The detailed results for this variant will be examined in subsection 8.3.2 after isolating some general trends from figure 8.1.

Prediction success increases in all cases with increasing dimensionality, which is to be expected as the data becomes more and more fully described. Note however that 100% success cannot be expected even using the full 1-755 m/z range due to the variable nature of mass spectra. The use of logarithmic spectra with and without a deadzone gives in each case better classification than does the binary form. The application of a deadzone markedly improves classification efficacy but at the expense of a high portion of unclassified spectra. This is illustrated in figure 8.2 where the average percentage of spectra for which a classification was not attempted, i.e. which lay within the deadzone, is plotted as a function of average dimensionality for both binary and logarithmic spectra for each of the training (Tr76) and the two prediction (Pr20 and Pr49) sets. The size of the deadzone was determined on the training set and then applied for prediction purposes, and as can be seen from figure 8.2 in each case more spectra from the prediction sets lay within the region of non classification than could be anticipated from the behaviour of the training set. The binary spectra were especially inconsistent in this respect, with 50-55% non classification on the prediction sets with an average of 61.2 mass positions and 30-35% with an average of 18.6 mass positions. This behaviour renders the binary data form invalid for such distance from the mean calculations on this data base.

8.3.2 Best Classification The best classification results using logarithmic spectra with no deadzone and an average of 61.2 m/z positions are presented in tables 8.3-8.5 for the training and first and second prediction sets. These results are summarised graphically by the measures figure of merit and overall success rate P_{tot} in figure 8.3. As can be seen from these histograms this classifier again performs better for recognition on the training set (Tr76) than on either of the prediction sets (Pr20 and Pr49), as

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(1|J)	P(2|N)	I(A,B)	FIG MER	IMAX
1	CT11	47	.618	.894	.793	.855	0.237	.875	.821	.366	.382	.380
2	CT12	35	.461	.943	.976	.961	0.421	.971	.952	.757	.761	.759
3	CT15	28	.368	.929	.854	.882	0.250	.788	.953	.472	.497	.511
4	OT5	43	.566	.767	.909	.829	0.263	.917	.750	.364	.369	.375
5	OT6	25	.329	.760	.941	.882	0.211	.864	.889	.390	.427	.417
6	C6	35	.461	.943	.902	.921	0.382	.892	.949	.605	.608	.610
7	C7	31	.408	.968	.844	.895	0.303	.811	.974	.546	.560	.574
8	C8	24	.316	.917	.962	.947	0.263	.917	.962	.608	.676	.674
9	C10	22	.289	.909	.963	.947	0.237	.909	.963	.578	.666	.664
10	O1	39	.513	.846	.973	.908	0.395	.971	.857	.587	.587	.589
11	O2	24	.316	.875	.923	.908	0.224	.840	.941	.475	.527	.531
12	N4	53	.697	.999	.870	.961	0.263	.946	.999	.662	.749	.708
13	N5	43	.566	.999	.848	.934	0.368	.896	.999	.683	.692	.677
14	NC6	45	.592	.889	.935	.908	0.316	.952	.853	.553	.567	.574
15	OC2	19	.250	.947	.965	.961	0.211	.900	.982	.593	.730	.741
16	PUR	46	.605	.978	.700	.868	0.263	.833	.955	.429	.443	.427
17	PYR	16	.211	.999	.967	.974	0.184	.889	.999	.623	.839	.894
18	ADN	45	.592	.956	.774	.882	0.289	.860	.923	.457	.469	.460
19	AN6	33	.434	.909	.930	.921	0.355	.909	.930	.590	.598	.597
20	ASUG	34	.447	.971	.762	.855	0.303	.767	.970	.464	.468	.477
21	S133	24	.316	.833	.981	.934	0.250	.952	.927	.551	.613	.591
	AV.			.916	.894	.911	0.285	.889	.931	.541	.582	.582

Table 8.3: Distance from mean analysis on Tr76. Logarithmic data used with zero deadzone and an average of 61.2 m/z values. For column headings see subsection 6.3.2.

Table 8.4: [Overleaf] Distance from mean analysis on Pr20. Logarithmic data used with zero deadzone and an average of 61.2 m/z values. For column headings see subsection 6.3.2.

Table 8.5: [Overleaf] Distance from mean analysis on Pr49. Logarithmic data used with zero deadzone and an average of 61.2 m/z values. For column headings see subsection 6.3.2.

Table 8.4

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(2|N) P(1|J)		I(A,B)	FIG MER	IMAX
1	CT11	13	.650	.923	.714	.850	0.200	.857	.833	.325	.348	.341
2	CT12	9	.450	.778	.999	.900	0.350	.999	.846	.590	.594	.582
3	CT15	8	.400	.875	.999	.950	0.350	.999	.923	.717	.738	.717
4	OT5	10	.500	.999	.700	.850	0.350	.769	.999	.493	.493	.493
5	OT6	7	.350	.857	.923	.900	0.250	.857	.923	.473	.506	.505
6	C6	9	.450	.889	.999	.950	0.400	.999	.917	.744	.750	.739
7	C7	8	.400	.750	.917	.850	0.250	.857	.846	.361	.372	.367
8	C8	7	.350	.857	.999	.950	0.300	.999	.929	.674	.722	.689
9	C10	7	.350	.857	.999	.950	0.300	.999	.929	.674	.722	.689
10	O1	10	.500	.999	.800	.900	0.400	.833	.999	.610	.610	.610
11	O2	6	.300	.500	.857	.750	0.050	.600	.800	.097	.110	.110
12	N4	14	.700	.929	.999	.950	0.250	.999	.857	.674	.765	.811
13	N5	11	.550	.909	.889	.900	0.350	.909	.889	.525	.528	.528
14	NC6	12	.600	.833	.625	.750	0.150	.769	.714	.162	.167	.166
15	OC2	5	.250	.800	.999	.950	0.200	.999	.938	.541	.667	.610
16	PUR	12	.600	.999	.875	.950	0.350	.923	.999	.717	.738	.717
17	PYR	4	.200	.500	.938	.850	0.050	.667	.882	.140	.194	.189
18	ADN	13	.650	.999	.999	.999	0.350	.999	.999	.934	.999	.999
19	AN6	9	.450	.778	.818	.800	0.250	.778	.818	.273	.275	.275
20	ASUG	7	.350	.999	.615	.750	0.100	.583	.999	.346	.371	.410
21	S133	8	.400	.375	.999	.750	0.150	.999	.706	.228	.235	.219
	AV.			.829	.889	.879	0.257	.876	.893	.490	.519	.513

Table 8.5

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(2|N) P(1|J)		I(A,B)	FIG MER	IMAX
1	CT11	34	.694	.765	.667	.735	0.041	.839	.556	.121	.137	.140
2	CT12	24	.490	.750	.760	.755	0.245	.750	.760	.197	.197	.197
3	CT15	15	.306	.733	.676	.694	0.000	.500	.852	.106	.120	.125
4	OT5	27	.551	.852	.636	.755	0.204	.742	.778	.191	.192	.191
5	OT6	14	.286	.714	.886	.837	0.122	.714	.886	.250	.290	.291
6	C6	22	.449	.773	.741	.755	0.204	.708	.800	.198	.199	.200
7	C7	15	.306	.667	.676	.673	-.020	.476	.821	.074	.083	.087
8	C8	14	.286	.643	.800	.755	0.041	.563	.848	.127	.147	.151
9	C10	12	.245	.667	.757	.735	-.020	.471	.875	.102	.127	.135
10	O1	27	.551	.815	.864	.837	0.286	.880	.792	.361	.364	.365
11	O2	11	.224	.636	.868	.816	0.041	.583	.892	.155	.202	.207
12	N4	37	.755	.919	.999	.939	0.184	.999	.800	.582	.725	.792
13	N5	30	.612	.833	.842	.837	0.224	.893	.762	.343	.356	.360
14	NC6	33	.673	.758	.688	.735	0.061	.833	.579	.133	.146	.149
15	OC2	12	.245	.833	.999	.959	0.204	.999	.949	.571	.711	.655
16	PUR	28	.571	.929	.714	.837	0.265	.813	.882	.349	.354	.349
17	PYR	10	.204	.800	.974	.939	0.143	.889	.950	.404	.553	.531
18	ADN	29	.592	.931	.800	.878	0.286	.871	.889	.440	.451	.446
19	AN6	17	.347	.706	.688	.694	0.041	.545	.815	.104	.112	.115
20	ASUG	17	.347	.824	.656	.714	0.061	.560	.875	.160	.172	.179
21	S133	16	.327	.563	.788	.714	0.041	.563	.788	.086	.095	.096
	AV.			.767	.785	.790	0.126	.723	.817	.241	.273	.274

is to be expected, although performance on Pr20 is only slightly below that on Tr76. The average figure of merit for Pr20 is 0.519 (table 8.4) as opposed to 0.582 (table 8.3) for Tr76. As was also evident in chapters 6 and 7 there is the same drop in performance between the original (Pr20) and the augmented (Pr49) prediction sets, with the average figure of merit decreasing to 0.273 (table 8.5) for the latter. These trends are also evident in the other eleven variants of the method the

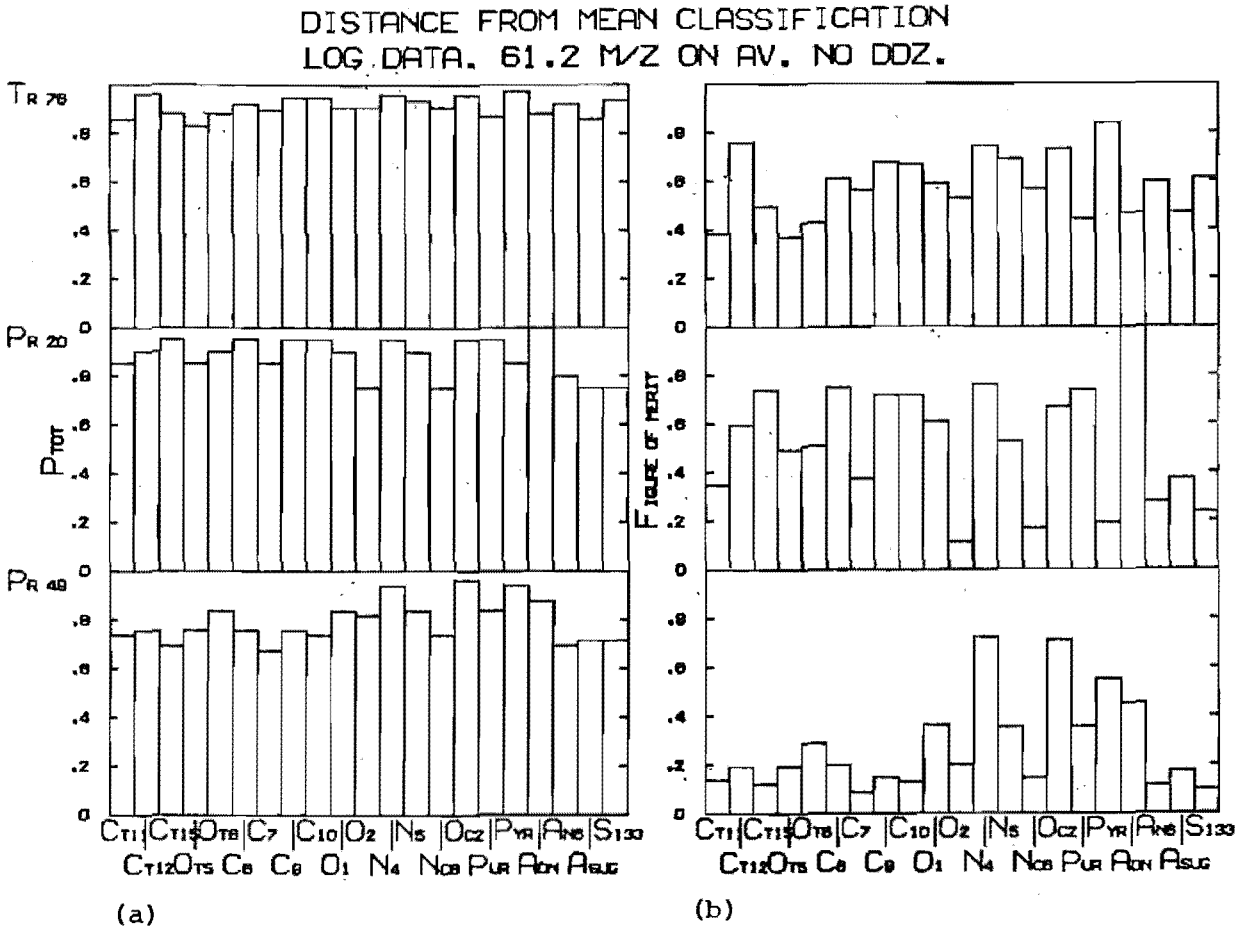


Figure 8.3: Histograms of (a) P_{tot} and (b) figure of merit for distance from mean classification. Logarithmic data used with zero deadzone and an average of 61.2 mass positions. Results depicted for training and prediction sets.

detailed results of which are not presented. In common with the forms of analysis discussed in chapters 6 and 7 it is not possible to rationalise differences between the twenty-one individual analyses of figure 8.3.

This particular variant of the distance from the mean classification technique compares moderately well with a similar distance from the mean classification of Rotter and Varmuza [145] on a set of 524 steroid mass spectra which gave, for seventeen structural classes, an average figure of merit value of 0.35. Wilkins et al. in their more sophisticated linear learning machine and sequential simplex optimisation work [141] using 60 component vectors and a much larger (1252 spectra) data base, obtained for eleven structural categories an average figure of merit of 0.42. The better results of these two groups of researchers may indicate that the smallness of the training set is again (cf. subsection 6.4.3) a reason for the lower performance of the classifiers presented here.

Chapter 9

NEAREST NEIGHBOUR APPROACH

9.1 Introduction

The k-nearest neighbour method involves the comparison of an unknown with each of the members of a training set of spectra in a selected d-dimensional hyperspace. The spectrum of the training set to which the unknown is closest determines its class. A variation is to consider several closest neighbours and allow each one vote in the assignment of the unknown. A theoretical statistical foundation of the method has been published [150]. The method has previously been applied to NMR spectra [150] and to Kovats' retention indices of GC liquid phases [19] as well as to mass spectra [143]. In the limit as the training set becomes large the method approximates to a library search (cf. subsection 2.3.2). In this present work the method has been applied to the data base described in subsection 5.2.1. The results obtained (subsection 9.3.2) are the poorest of any of the pattern recognition techniques used in this work, and the computation involved was the most expensive (section 9.2). The method is seen as being unsuitable for this data base and probably for nucleoside spectra in general.

9.2 Method

The k-nearest neighbour approach is based upon the truism that the best description of the data is the data itself [143]. The simplest case is to assign an unknown j as its single (k=1) nearest neighbour i of the training set Tr76, using the generalised euclidean distance

$$D_{ij} = \sqrt{\sum_{\ell=1}^{d} (x_{i\ell} - x_{j\ell})^2} \quad . \tag{9.2.1}$$

The cases with k = 3,5 and 7 nearest neighbours were also investigated. Each neighbour is given one vote V_i, +1 for class members and -1 for class non members, and the sum of the votes

$$\sum_{i=1}^{k} V_i \qquad k = 3,5,7 \tag{9.2.2}$$

determines the assignment of the unknown. Two weighting techniques were investigated following the suggestion of Jurs and Isenhour [145], division of each vote by the distance D_{ij} and the square of the distance

$$\sum_{i=1}^{k} V_i/D_{ij} \quad , \quad \sum_{i=1}^{k} V_i/D_{ij}^2 \qquad k = 3,5,7 \tag{9.2.3}$$

from the unknown j.

The computational approach adopted was to set up a distance matrix between each of the 76 spectra in the training set for recognition purposes, and between the 76 training spectra and the 49 prediction spectra for prediction on unknowns. These matrices were scanned to find the 1st, 2nd, ... 7th nearest neighbours of and the distances from each spectrum to be classified, and the values for the various k and weighting schemes were obtained accordingly. The analyses were conducted with d = 82, 24 and 8-14 dimensions, i.e. m/z positions, using the m/z selections obtained as described in section 6.2. The spectra were presented to the classifiers in the forms described in subsection 5.2.2 i.e. autoscaled logarithmic and binary spectra with prior normalisation to 100% of the base peak and with retention of only peaks $\geqslant$ 1% relative intensity and $\geqslant$ 100 amu.

The most time consuming part of the analysis was establishment of the distance matrix and CPU times for this are shown in table 9.1. These are for logarithmic data while with binary data the process took in each case 1-2 sec less. Once the distance matrix had been established selection of the 1,3,5 or 7 nearest neighbours and weights always took under 5 sec. The 82 m/z variant was thus computationally the most expensive of any used in this work. The k-nearest neighbour approach together with the distance from the mean method of the previous chapter

	Recognition	Prediction
82 m/z	36-41	37-42
24 m/z	17-22	15-19
8-14 m/z	11-15	8-12

Table 9.1: CPU times for k-nearest neighbour analyses. Times in seconds shown for recognition on TR76 and prediction on Pr49 with logarithmic data.

were encoded in program KNNCLASSIF which is reproduced in appendix II.

9.3 Results and Discussion

9.3.1 Variants of Method Seventy-two variations on the k-nearest neighbour approach were applied to the twenty-one structural features described in subsection 5.2.3. A number of these variants returned identical or nearly identical classifications and not all are reported in detail here. Average prediction success on the twenty-one categories is shown in table 9.2 for those twenty-four variants using simple sums of votes i.e. without distance weighting. The twenty-four variants were, in summary, combinations of logarithmic/binary data, k = 1,3,5 and 7 nearest

(Pr49)		82 m/z		24 m/z		8-14 m/z	
	$\sum_i V$	M	P_{tot}	M	P_{tot}	M	P_{tot}
log	k=1	.080	62.9%	.082	65.6%	.074	64.2%
	k=3	.065	63.3	.073	64.4	.060	63.9
	k=5	.049	61.3	.053	63.8	.044	62.5
	k=7	.031	58.7	.027	61.5	.028	61.0
bin	k=1	.125	61.4%	.083	59.3%	.069	54.5%
	k=3	.073	57.9	.092	61.1	.073	57.1
	k=5	.064	58.0	.062	59.6	.066	59.0
	k=7	.066	60.0	.065	61.3	.061	59.7

Table 9.2: k-nearest neighbour classification success on Pr49. Average figure of merit M and P_{tot} values for binary and autoscaled logarithmic data with three sets of m/z values as described in section 9.2. Sum of votes variant of method used with k = 1,3,5 and 7 nearest neighbours.

neighbours, and 82/24/8-14 mass positions. These are graphed as functions of numbers of nearest neighbours in figure 9.1 using the average figure of merit on the twenty-one categories as the criterion

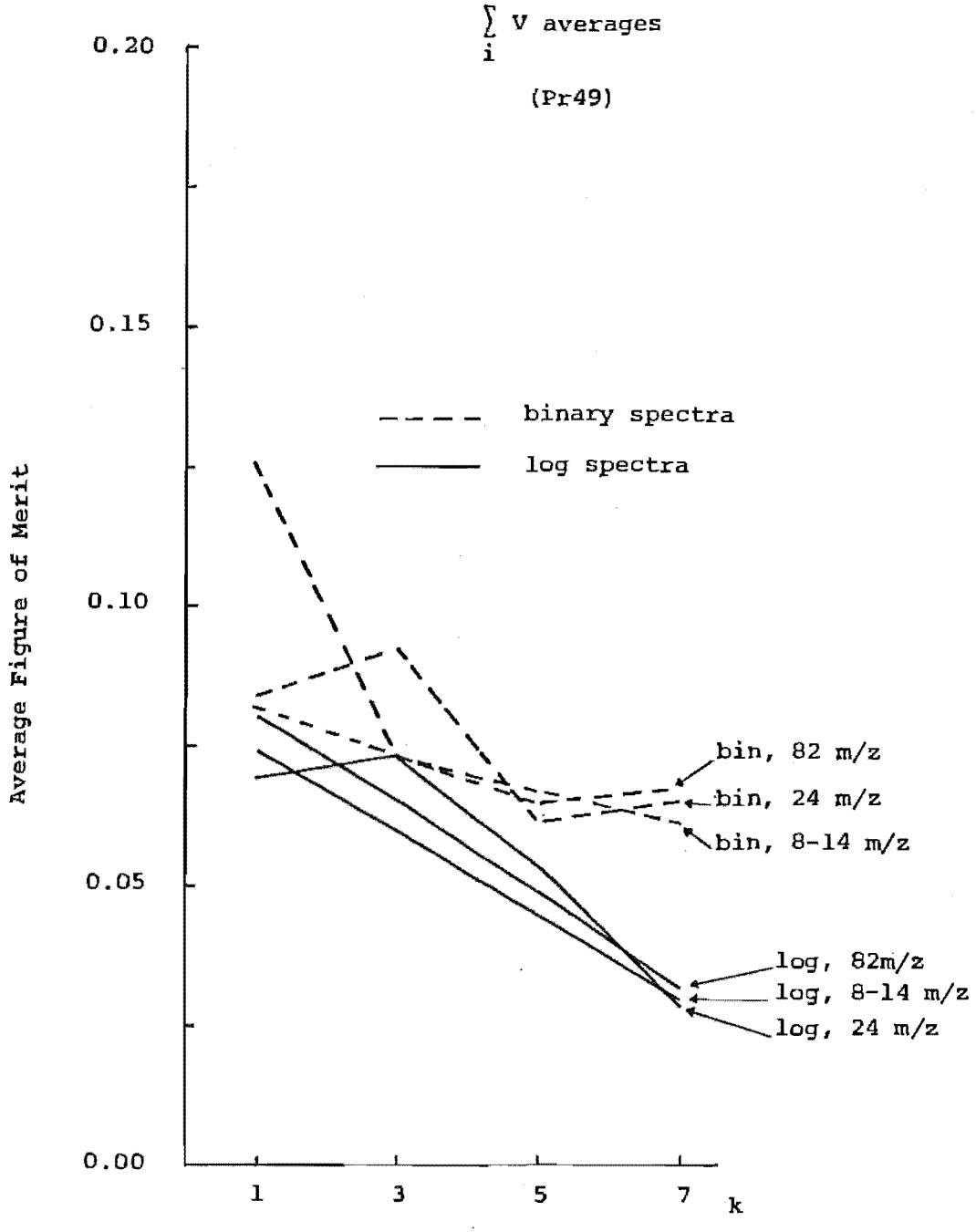


Figure 9.1: Graph of average k-nearest neighbour classification success. Average figure of merit plotted against number of nearest neighbours, for binary and logarithmic data using three m/z selections. Results for augmented prediction set Pr49.

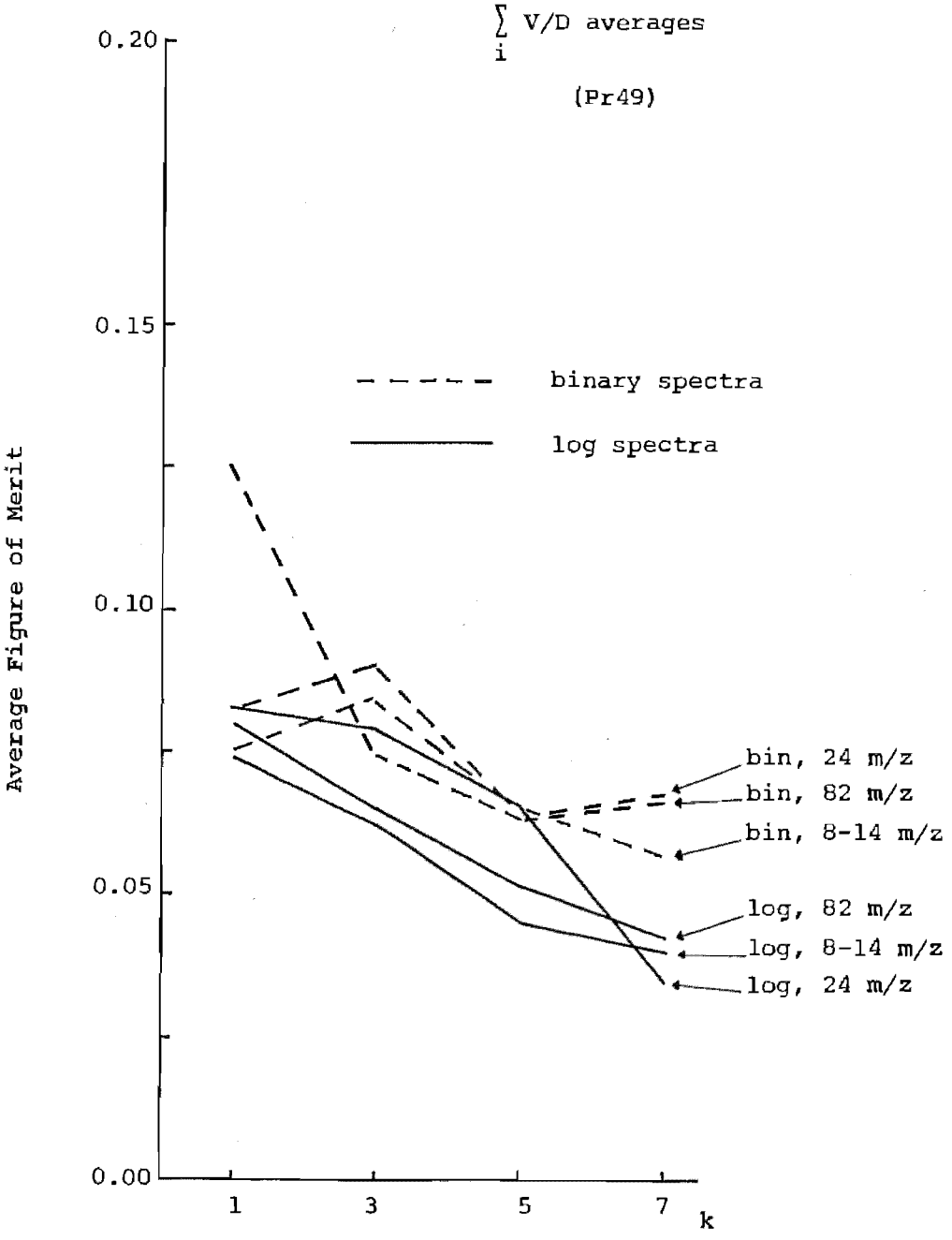


Figure 9.2: Graphs of average k-nearest neighbour classification success using $\sum_i$ V/D. Average figure of merit plotted against number of nearest neighbours, for binary and logarithmic data using three m/z selections. Results for augmented prediction set Pr49.

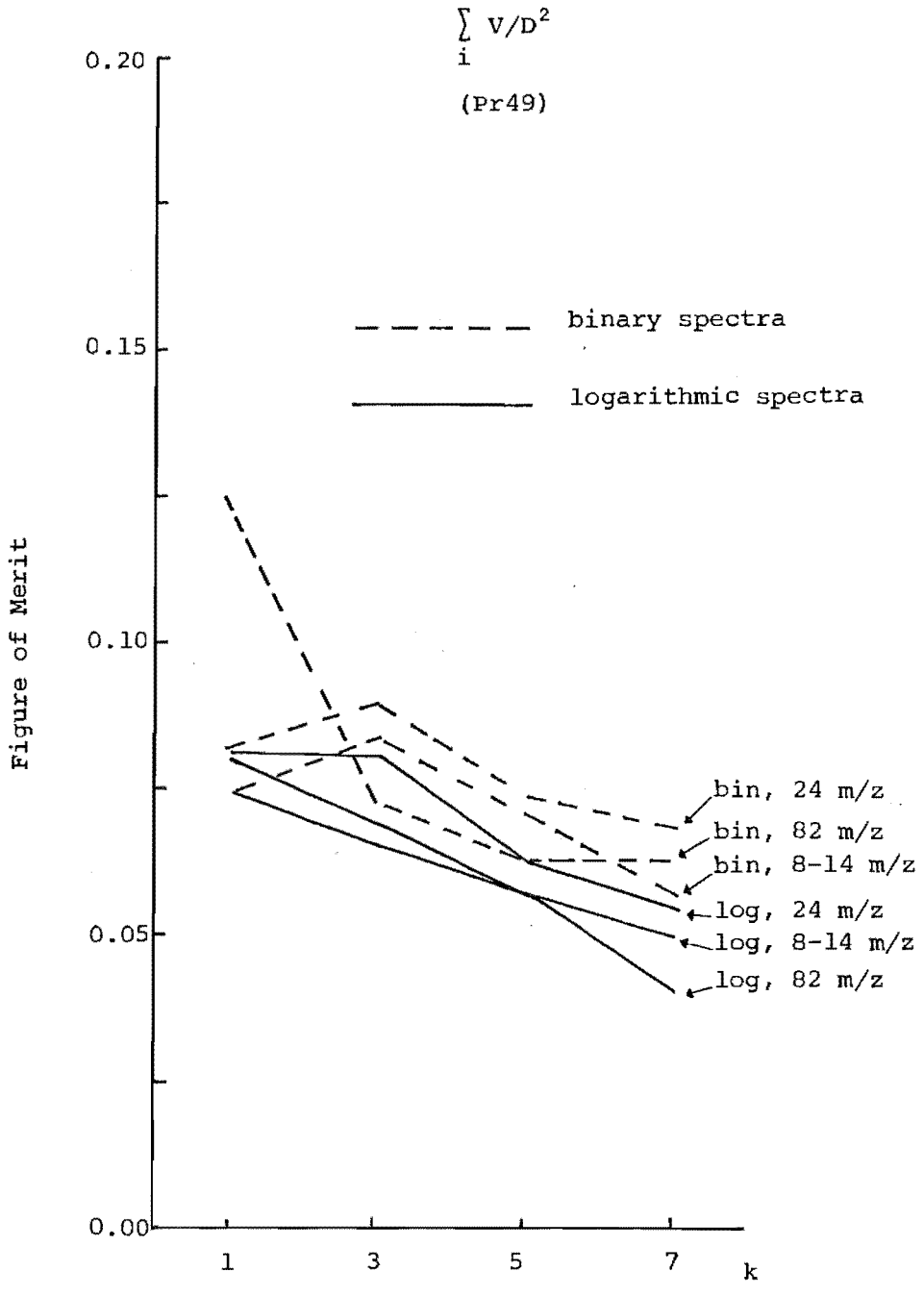


Figure 9.3: Graph of average k-nearest neighbour classification success using $\sum_i V/D^2$. Average figure of merit plotted against number of nearest neighbours, for binary and logarithmic data using three m/z selections. Results for augmented prediction set Pr49.

of "goodness" of prediction. The less precise but more intuitively obvious measure of percentage overall prediction success, P_{tot}, is presented in table 9.2 as well as the figure of merit but is not used for comparison of the methods. A critique of these two performance evaluation measures is contained in subsection 9.3.3. Analogous results for the two distance weighted sums of votes differ very little from the unweighted case, and consequently are presented in graphical form only in figure 9.2 for $\Sigma V/D$ and figure 9.3 for $\Sigma V/D^2$. Only the averages for the augmented prediction set Pr49 are recorded for most of the seventy-two variants as much of the discussion will centre on these.

As can be seen from figure 9.1 or table 9.2 the classifications achieved using any variant of this method are rather poor, although comparison with similar studies by other workers is delayed until subsection 9.3.2 and with other methods in this present work until chapter 10. The best average prediction is achieved, as can be seen from figure 9.1, with binary spectra using 82 mass positions and k=1 nearest neighbour, which variant gives $M_{av.} = 0.125$ (table 9.2). A detailed analysis of this variant is given in subsection 9.3.2.

This is a surprising result in view of the fact that maximum information should be retained by the logarithmic form, the average figure of merit for which is only 0.080 (table 9.2) for an equivalent number of mass positions. In fact the binary form invariably performs better for all k and all dimensionalities than does the logarithmic, as is obvious from figure 9.1, and this phenomenon is carried through to the distance weighted cases of figures 9.2 and 9.3. A possible explanation for this superiority of the binary data form with this method is the significant variation of intensity in nucleoside mass spectra, and the reliance upon intensities of individual spectra rather than upon a value smoothed by averaging over a set of twenty or thirty class members. Thus while an averaged intensity as in the distance from the mean approach of chapter 8 may give reasonable comparison with all spectra, the intensities of any two individual spectra will often show wide variation. The reduction to binary form may thus act in this context as a form of data smoothing, reducing individual variations in intensity and hence the variation of distance between pattern vectors belonging to the same category. Such a phenomenon, if this is indeed the correct explanation, has not been previously reported.

The other important trend from figure 9.1 or, equivalently, from figures 9.2 or 9.3 is the decrease in predictive ability with increasing number k of neighbours. For all except the 24 m/z variants, both binary and logarithmic, k = 1 gives the best classification. This has previously been noticed in other nearest neighbour studies by Kowalski and Bender [150]. Comparison of figure 9.1 with the two distance weighted cases of figures 9.2 and 9.3 reveals very little increase or even variation in predictive ability in the latter two variants. The three sets of graphs for at least k = 1,3,5 are almost identical, thus rendering the extra computation pointless.

9.3.2 Best Classification The most efficacious variant of the method, as noted above, is binary data with 82 m/z positions and k=1, using the unweighted sum of votes. The detailed results for each of the twenty-one categories are presented in table 9.3 for the training set and tables 9.4 and 9.5 for the first and second prediction sets. The data of the three tables is presented graphically in figure 9.4 where histograms of P_{tot} and figure of merit have been drawn for the three sets. Almost 100% recognition is achieved on the training set (table 9.3, graphed in the top histograms of figure 9.4) as is inherent in the truism that a pattern's nearest neighbour is itself. The sole exception to perfect recognition is the fifth category, OT6, in which two compounds differing slightly in oxygen content have spectra with peaks at exactly the same mass positions of the 82 chosen. It should be pointed out that the analogous logarithmic variant gives 100% recognition on all categories of the training set, including OT6.

Prediction on Pr20 and Pr49 is of course much less than recognition on Tr76, with average figures of merit 0.273 and 0.125 respectively (tables 9.4 and 9.5). There is the same drop in performance between the two prediction sets evident either from these two averages or by comparison of the central and lower histograms of figure 9.4 as has been noted before (subsection 6.4.3). The structural category N4, base nitrogen number $\geqslant 4$, appears from figure 9.4(b) to be the best classified by this method. Again, differences in prediction between individual analyses are not rationalised.

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(2|N) P(1|J)		I(A,B)	FIG MER	IMAX
1	CT11	47	.618	.999	.999	.999	0.382	.999	.999	.959	.999	.999
2	CT12	35	.461	.999	.999	.999	0.461	.999	.999	.995	.999	.999
3	CT15	28	.368	.999	.999	.999	0.368	.999	.999	.949	.999	.999
4	OT5	43	.566	.999	.999	.999	0.434	.999	.999	.987	.999	.999
5	OT6	25	.329	.999	.980	.987	0.316	.962	.999	.833	.912	.930
6	C6	35	.461	.999	.999	.999	0.461	.999	.999	.995	.999	.999
7	C7	31	.408	.999	.999	.999	0.408	.999	.999	.975	.999	.999
8	C8	24	.316	.999	.999	.999	0.316	.999	.999	.900	.999	.999
9	C10	22	.289	.999	.999	.999	0.289	.999	.999	.868	.999	.999
10	O1	39	.513	.999	.999	.999	0.487	.999	.999	.999	.999	.999
11	O2	24	.316	.999	.999	.999	0.316	.999	.999	.900	.999	.999
12	N4	53	.697	.999	.999	.999	0.303	.999	.999	.884	.999	.999
13	N5	43	.566	.999	.999	.999	0.434	.999	.999	.987	.999	.999
14	NC6	45	.592	.999	.999	.999	0.408	.999	.999	.975	.999	.999
15	OC2	19	.250	.999	.999	.999	0.250	.999	.999	.811	.999	.999
16	PUR	46	.605	.999	.999	.999	0.395	.999	.999	.968	.999	.999
17	PYR	16	.211	.999	.999	.999	0.211	.999	.999	.742	.999	.999
18	ADN	45	.592	.999	.999	.999	0.408	.999	.999	.975	.999	.999
19	AN6	33	.434	.999	.999	.999	0.434	.999	.999	.987	.999	.999
20	ASUG	34	.447	.999	.999	.999	0.447	.999	.999	.992	.999	.999
21	S133	24	.316	.999	.999	.999	0.316	.999	.999	.900	.999	.999
	AV.			.999	.999	.999	0.373	.998	.999	.933	.996	.997

Table 9.3: k-nearest neighbour analysis on Tr76. Binary data used with 82 m/z values and k=1, using the unweighted sum of votes. For column headings see subsection 6.3.2.

Table 9.4: [Overleaf] k-nearest neighbour analysis on Pr20. Binary data used with 82 m/z values and k=1, using the unweighted sum of votes. For column headings see subsection 6.3.2.

Table 9.5: [Overleaf] k-nearest neighbour analysis on Pr49. Binary data used with 82 m/z values and k=1, using the unweighted sum of votes. For column headings see subsection 6.3.2.

Table 9.4

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(1|J)	P(2|N) I(A,B)	FIG MER	IMAX
1	CT11	13	.650	.769	.857	.800	0.150	.909	.667	.279	.299	.309
2	CT12	9	.450	.889	.545	.700	0.150	.615	.857	.161	.162	.165
3	CT15	8	.400	.625	.999	.850	0.250	.999	.800	.430	.442	.419
4	OT5	10	.500	.999	.400	.700	0.200	.625	.999	.236	.236	.236
5	OT6	7	.350	.857	.538	.650	0.000	.500	.875	.117	.125	.132
6	C6	9	.450	.778	.545	.650	0.100	.583	.750	.080	.081	.082
7	C7	8	.400	.999	.583	.750	0.150	.615	.999	.346	.357	.381
8	C8	7	.350	.857	.846	.850	0.200	.750	.917	.361	.387	.394
9	C10	7	.350	.999	.769	.850	0.200	.700	.999	.493	.528	.572
10	O1	10	.500	.900	.700	.800	0.300	.750	.875	.296	.296	.296
11	O2	6	.300	.667	.000	.200	-.500	.222	.000	.000	.000	.000
12	N4	14	.700	.999	.999	.999	0.300	.999	.999	.881	.999	.999
13	N5	11	.550	.909	.556	.750	0.200	.714	.833	.194	.195	.193
14	NC6	12	.600	.000	.999	.400	-.200	.000	.400	.000	.000	.000
15	OC2	5	.250	.999	.533	.650	-.100	.417	.999	.223	.275	.338
16	PUR	12	.600	.667	.750	.700	0.100	.800	.600	.125	.128	.130
17	PYR	4	.200	.750	.563	.600	-.200	.300	.900	.047	.065	.075
18	ADN	13	.650	.923	.857	.900	0.250	.923	.857	.473	.506	.505
19	AN6	9	.450	.999	.273	.600	0.050	.529	.999	.145	.146	.152
20	ASUG	7	.350	.999	.462	.650	0.000	.500	.999	.234	.251	.281
21	S133	8	.400	.750	.833	.800	0.200	.750	.833	.256	.264	.264
	AV.			.826	.648	.707	0.086	.629	.817	.256	.273	.282

Table 9.5

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(1|J)	P(2|N) I(A,B)	FIG MER	IMAX
1	CT11	34	.694	.647	.667	.653	-.041	.815	.455	.061	.069	.072
2	CT12	24	.490	.667	.440	.551	0.041	.533	.579	.009	.009	.009
3	CT15	15	.306	.667	.647	.653	-.041	.455	.815	.061	.069	.072
4	OT5	27	.551	.852	.273	.592	0.041	.590	.600	.017	.017	.017
5	OT6	14	.286	.714	.400	.490	-.224	.323	.778	.008	.010	.010
6	C6	22	.449	.773	.519	.633	0.082	.567	.737	.066	.066	.067
7	C7	15	.306	.999	.500	.653	-.041	.469	.999	.237	.267	.311
8	C8	14	.286	.786	.514	.592	-.122	.393	.857	.057	.066	.072
9	C10	12	.245	.917	.568	.653	-.102	.407	.955	.146	.182	.210
10	O1	27	.551	.852	.409	.653	0.102	.639	.692	.063	.063	.063
11	O2	11	.224	.727	.211	.327	-.449	.211	.727	.000	.000	.000
12	N4	37	.755	.999	.833	.959	0.204	.949	.999	.571	.711	.655
13	N5	30	.612	.967	.579	.816	0.204	.784	.917	.293	.304	.292
14	NC6	33	.673	.000	.999	.327	-.347	.000	.327	.000	.000	.000
15	OC2	12	.245	.667	.405	.469	-.286	.267	.789	.003	.004	.004
16	PUR	28	.571	.750	.619	.694	0.122	.724	.650	.101	.103	.103
17	PYR	10	.204	.800	.487	.551	-.245	.286	.905	.042	.058	.067
18	ADN	29	.592	.897	.650	.796	0.204	.788	.813	.246	.252	.249
19	AN6	17	.347	.999	.313	.551	-.102	.436	.999	.145	.156	.177
20	ASUG	17	.347	.824	.500	.612	-.041	.467	.842	.077	.083	.087
21	S133	16	.327	.813	.606	.673	0.000	.500	.870	.119	.130	.137
	AV.			.777	.530	.614	-.050	.505	.776	.111	.125	.127

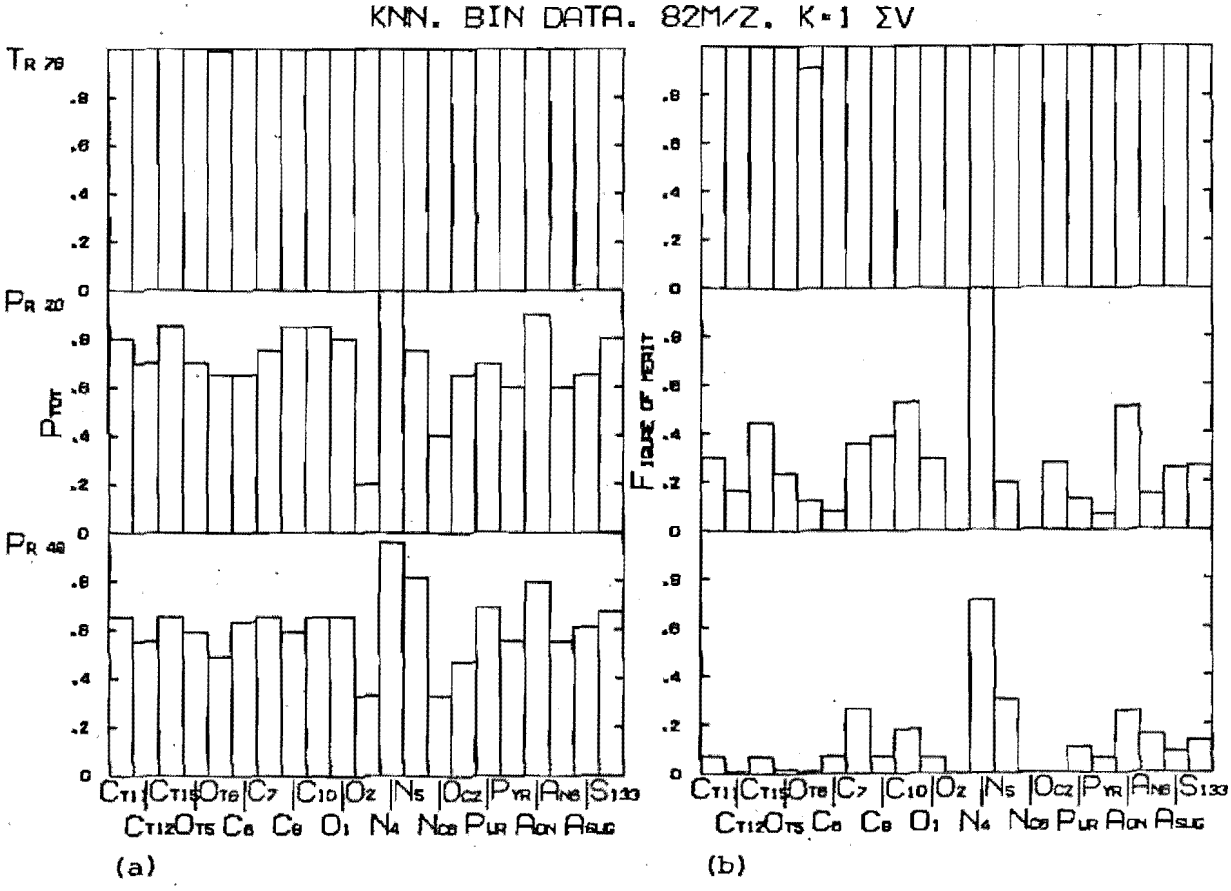


Figure 9.4: Histograms of (a) P_{tot} and (b) figure of merit for k-nearest neighbour classification. Binary data used with 82 m/z values and k=1, using the unweighted sum of votes. Results depicted for training and prediction sets.

Similar studies have achieved far better classifications than are attained here. The k-nearest neighbour method is the worst of the pattern recognition techniques used in this work and this point will be explored in more depth in the next chapter. The average figure of merit values of Wilkins et al. [141] of 0.42 or of Rotter and Varmuza [145] of 0.35 far exceed the 0.125 reported here. The k-nearest neighbour approach therefore appears to be unsuitable for classification of this data base, and the poor performance strongly indicates that the method is unsuitable for nucleoside mass spectra in general. On the augmented prediction set Pr49 the average success rate, P_{tot}, is

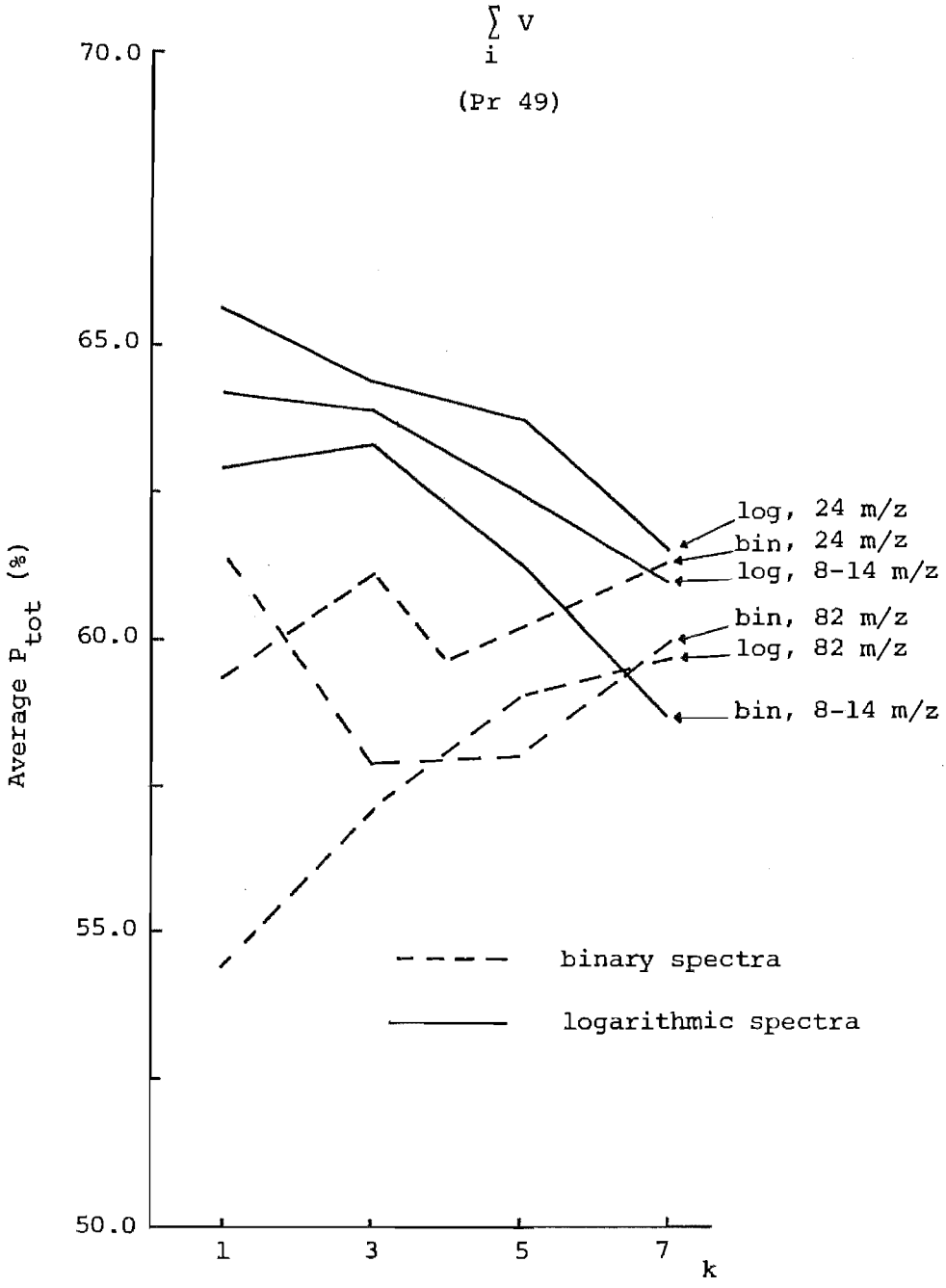


Figure 9.5: Graph of k-nearest neighbour classification success. Average percentage success rate P_{tot} plotted against number of nearest neighbours, for binary and logarithmic data using three m/z selections. Results for augmented prediction set Pr49.

only 61.4% (table 9.5), and the method actually performs worse than simple assignment to the more populous class ("improv most pop" column of table 9.5 averages -5.0%).

9.3.3 Classifier Evaluation Although not uniquely related to the k-nearest neighbour method, the opportunity is taken here to present practical examples of some of the points first raised in subsections 5.3.1 and 5.3.2. The results achieved by this method illustrate the greater preciseness of the figure of merit concept as opposed to the simple percentage success rate P_{tot}. Inspection of table 9.2 reveals that although the binary form with k=1 and 82 m/z positions gives a much greater average M value than does the logarithmic (0.125 against 0.080), the P_{tot} averages for the same two variants are nearly equal, with that of the logarithmic form being actually slightly greater (62.9% against 61.4%). In fact if the graphs of figure 9.1 are replotted but using the average P_{tot} criterion as in figure 9.5, a very different picture emerges.

For the logarithmic forms the positions of the 82 m/z and the 8-14 m/z lines reverse, while the 24 m/z variant is superior for all k instead of only k = 3 and 5 as was the case using the figure of merit. The binary and logarithmic forms exchange positions, making the latter appear clearly superior, and the former reverse the trend of decreasing prediction with increasing number k of neighbours. Using this criterion the best variant is now logarithmic data with k = 1 and 24 mass positions, with average P_{tot} = 65.6% corresponding to $M_{av.}$ = 0.082 (table 9.2). Prediction using this logarithmic form on Pr49 is tabulated in table 9.6.

Comparison of table 9.6 and table 9.5 reveals that the average success rates for the class members and non members separately, P_1 and P_2, are 76.4% and 50.0% (table 9.6) for the logarithmic 24 m/z form, and 77.7% and 53.0% (table 9.5) for the binary 82 m/z. On this basis the latter is a better classifier although the difference is very small. The magnitude of the corresponding change in the average figure of merit from 0.125 to 0.082 is a reflection of the sensitivity of this measure to small variations in predictive ability.

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(1|J)	P(2|N)	I(A,B)	FIG MER	IMAX
1	CT11	34	.694	.853	.467	.735	0.041	.784	.583	.080	.090	.090
2	CT12	24	.490	.750	.520	.633	0.122	.600	.684	.056	.056	.056
3	CT15	15	.306	.533	.735	.673	-.020	.471	.781	.048	.054	.055
4	OT5	27	.551	.630	.636	.633	0.082	.680	.583	.051	.052	.052
5	OT6	14	.286	.786	.686	.714	0.000	.500	.889	.137	.159	.169
6	C6	22	.449	.864	.519	.673	0.122	.594	.824	.123	.124	.125
7	C7	15	.306	.533	.559	.551	-.143	.348	.731	.005	.006	.006
8	C8	14	.286	.714	.771	.755	0.041	.556	.871	.148	.172	.178
9	C10	12	.245	.833	.757	.776	0.020	.526	.933	.200	.249	.271
10	O1	27	.551	.741	.636	.694	0.143	.714	.667	.106	.107	.106
11	O2	11	.224	.636	.816	.776	0.000	.500	.886	.116	.151	.159
12	N4	37	.755	.973	.083	.755	0.000	.766	.500	.009	.012	.011
13	N5	30	.612	.900	.158	.612	0.000	.628	.500	.005	.005	.005
14	NC6	33	.673	.909	.313	.714	0.041	.732	.625	.053	.059	.058
15	OC2	12	.245	.750	.757	.755	0.000	.500	.903	.146	.181	.194
16	PUR	28	.571	.857	.095	.531	-.041	.558	.333	.000	.000	.000
17	PYR	10	.204	.700	.718	.714	-.082	.389	.903	.086	.117	.130
18	ADN	29	.592	.931	.200	.633	0.041	.628	.667	.027	.028	.028
19	AN6	17	.347	.706	.250	.408	-.245	.333	.615	.000	.000	.000
20	ASUG	17	.347	.882	.031	.327	-.327	.326	.333	.000	.000	.000
21	S133	16	.327	.563	.788	.714	0.041	.563	.788	.086	.095	.096
	AV.			.764	.500	.656	-.008	.557	.695	.071	.082	.085

Table 9.6: k-nearest neighbour analysis on Pr49. Logarithmic data used with 24 m/z values and k=1, using the unweighted sum of votes. For column headings see subsection 6.3.2.

Chapter 10

CONCLUSION AND COMPARISON OF METHODS

In this final chapter the two main approaches, heuristic programming and pattern recognition, are compared in so far as this is possible. The various pattern recognition methods are compared amongst themselves.

10.1 Heuristic Programming and Pattern Recognition

These two approaches are fundamentally different with respect to concept, method, and even, in this present application, aims. The main heuristic programming sought to identify numerical values associated with an unknown nucleoside, viz. molecular and base formula weight. Quantities such as these are not in general amenable to pattern recognition analysis, and consequently comparison of the two methods is difficult. The third aim of program NUCL, identification of the nature of the base (section 4.4), is in theory resolvable by both approaches but in the present work practical obstacles were encountered, viz. (a) programmatic difficulties and data base inadequacies (subsection 4.4.2) for program NUCL, and (b) under-representation of certain compound categories in the data base thereby prohibiting pattern recognition analyses on most base types (subsection 5.2.3). In fact only the most common base, adenine, could be examined, and no comparison is possible with program NUCL.

While a direct comparison of the two approaches is not feasible, it can be pointed out that each approach has roughly comparable success in achieving its own particular aims. Thus the success of the best pattern recognition variant, distance from the mean with logarithmic data, at identifying structural category was 79% (see table 10.3 of the next section). Conversely, molecular weight was correctly identified in 86% of the cases by program MOLION, and in 76% by the losses from M routine of program NUCL (table 4.1). Base formula weight was correctly obtained by program NUCL with 85% success. The facts that all these success rates are considerably better than random guessing and that they lie in the range 76%-86% points to the efficiency of the respective programs. Conversely the fact that they are in each case considerably less than 100% highlights the opinion that nucleoside spectra cannot as yet be routinely identified by computer techniques.

10.2 Pattern Recognition Studies

The four pattern recognition methods of chapters 6,7,8 and 9 performed with varying success on the data base of 125 nucleoside mass spectra. A summary of the results of these four chapters is reproduced in table 10.1 where the figure of merit values on each of the twenty-one structural categories is listed, for that variant of each of the four methods which performed best. The results apply to the augmented prediction set of 49 spectra Pr49. The corresponding recognition performances on the training set of 76 spectra Tr76 are recorded in table 10.2. Both these tables are condensations of material already presented in chapters 6-9. The figure of merit values of table 10.1 are graphed for each of the four best variants in figure 10.1. To provide a more readily grasped evaluation measure, the corresponding P_{tot} values for each of the four methods on Tr76 and Pr49 are summarised in table 10.3 by averages only over the twenty-one classes.

As can be seen from the averages of table 10.1 the best prediction is provided by the distance from the mean method, with an average figure of merit of 0.273 as opposed to 0.237 for the learning machine approach and 0.207 for the statistical linear discriminant function analysis. The worst performance is easily given by the k-nearest neighbour method with an average figure of merit of only 0.125. The best prediction average corresponds to an average overall success rate of 79.0% and the worst (0.125) to 61.4% (table 10.3). These are respectively 12.6% better (table 8.5) and 5.0% worse (table 9.5) classifications than would be achieved by simply assigning every spectrum to the more populous group (class members or non members) within each structural category. The best value of 0.273 compares reasonably well with the steroid study of Rotter and Varmuza [145], which gave on seventeen structural categories an average figure of merit of 0.35. It compares somewhat less well with the work of Wilkins et al. [141] which on eleven structural categories of mono-functionals generally smaller than those used here (molecular weight $\lesssim$ 300 amu as opposed to the range of 211 $\leqslant$ molecular weight $\leqslant$ 755 amu for the compounds in this work) returned an average figure of merit of 0.42.

The lower performance is explicable in terms of four factors:

(a) the enforced smallness of the data base (one tenth the size of the 1252 set of Wilkins et al.),

Table 10.1: Best figure of merit values for pattern recognition methods. Augmented prediction set (Pr49) results summarised from previous tables as indicated. Av/11 values are averages over only those eleven categories which were shown by the learning machine approach to be linearly separable in the chosen pattern space.

(Pr49)	Stat. lin. discrim.[a]	lma[b]	Distmean ddz 0.0[c]	knn k=1,ΣV[d]
Dimens.	24	24	61.2(av)	82
Preproc.	log	log	log	bin
Source table	6.8	7.4	8.5	9.5
1 CT11	.099	-	.137	.069
2 CT12	.115	-	.197	.009
3 CT15	.223	.175	.120	.069
4 OT5	.017	-	.192	.017
5 OT6	.223	-	.290	.010
6 C6	.115	.132	.199	.066
7 C7	.083	.145	.083	.267
8 C8	.274	.183	.147	.066
9 C10	.143	.291	.127	.182
10 O1	.085	.019	.364	.063
11 O2	.461	.255	.202	.000
12 N4	.303	.670	.725	.711
13 N5	.290	-	.356	.304
14 NC6	.089	-	.146	.000
15 OC2	.590	.303	.711	.004
16 Pur	.100	-	.354	.103
17 Pyr	.865	.339	.553	.056
18 Adn	.144	-	.451	.252
19 AN6	.041	.094	.112	.156
20 Asug	.027	-	.172	.083
21 S133	.055	-	.095	.130
Av.	.207	.237	.273	.125
Av/11	.289	.237	.304	.149

[a] statistical linear discriminant function analysis

[b] learning machine approach

[c] distance from mean with zero deadzone

[d] k-nearest neighbour classification using k=1 and unweighted sum of votes

(b) the structural diversity and complexity of the nucleosides considered,

(c) the known variability of nucleoside mass spectra, and

(d) the specificity of the structural features examined, i.e. distinction between the different types of nucleosides rather than more general distinctions such as separating nucleosides, peptides, etc from other classes of natural products as could be done with a more broadly based set of spectra.

These factors made the classifications attempted a severe test of the methods employed, and the results achieved do not as yet encourage the routine on line analytical identification of nucleosides by pattern recognition.

The relative order of the four approaches is a little surprising. The k-nearest neighbour method gave the best results of a number of such simple techniques compared by Justice and Isenhour [143] whereas the poor prediction here and the lengthy computations involved make it clearly the worst of the four methods. The learning machine and distance from the mean approaches differ little with average figures of merit 0.237 and 0.273, but the statistically based linear discriminant function analysis gives surprisingly good classification and the method would seem to be statistically very robust. It is not generally possible to determine a priori which method will perform best on any given data base.

The better spectral pre-processing of the two forms investigated was clearly autoscaled logarithmic data, as a significant proportion of the information content appears to be lost on reduction to binary form. Except for the k-nearest neighbour method which with logarithmic data gave very poor classification indeed (average figure of merit 0.082 from table 9.6) the best pre-processing for the other three methods was consistent with previous work [223]. Classification improved with increasing dimensionality, as is only to be expected as the data becomes more and more fully described. For both the distance from the mean and the k-nearest neighbour approaches maximum classification was attained with 82 mass positions, while for the other two methods where the parameter fitting procedures required $n/d \geqslant 3$ (subsection 5.1.1) maximum classification was achieved with 24 mass positions. When the distance from the mean approach was used with only 24 m/z values however performance dropped to $M_{av.} = 0.192$ (table 8.2) leaving the linear learning machine as the best classifier in this dimension pattern space.

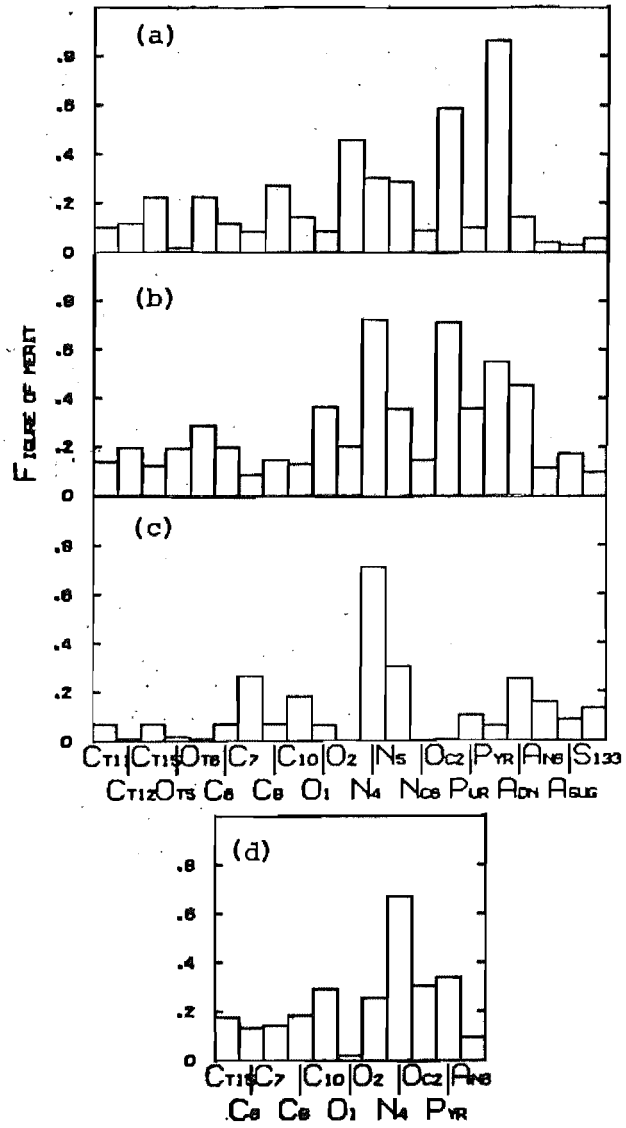


Figure 10.1: Histograms of pattern recognition prediction results. Figure of merit entries of table 10.2 for Pr49 using the methods of (a) statistical linear discriminant function analysis, (b) distance from the mean, (c) k-nearest neighbour, and (d) learning machine approach.

The twenty-one individual analyses show few consistent trends and performance for each category is almost entirely dependent upon the method employed, as is obvious from figure 10.1 or table 10.1. This is only to be expected and the sole exception seems to be N4 which is very well classified by all methods. It is in fact the category best classified by three of the methods (figure 10.1). This could perhaps be related

(Tr76)	Stat. lin. discrim.	lma	Distmean ddz 0.0	knn k-1,ΣV
Dimens.	24	24	61.2 (av.)	82
Preproc.	log	log	log	bin
Source Table	6.6	7.2	8.3	9.3
1 CT11	.692	-	.382	.999
2 CT12	.910	-	.761	.999
3 CT15	.904	.999	.497	.999
4 OT5	.628	-	.369	.999
5 OT6	.619	-	.427	.912
6 C6	.846	.999	.608	.999
7 C7	.999	.999	.560	.999
8 C8	.999	.999	.676	.999
9 C10	.805	.999	.666	.999
10 O1	.850	.999	.587	.999
11 O2	.895	.999	.527	.999
12 N4	.999	.999	.749	.999
13 N5	.758	-	.692	.999
14 NC6	.821	-	.567	.999
15 OC2	.907	.999	.730	.999
16 Pur	.695	-	.443	.999
17 Pyr	.999	.999	.839	.999
18 Adn	.648	-	.469	.999
19 AN6	.713	.999	.598	.999
20 Asug	.529	-	.468	.999
21 S133	.702	-	.613	.999
Av.	.806	.999	.582	.996
Av/11	.901	.999	.640	.999

Table 10.2: Best figure of merit values for pattern recognition methods, on training set. Entries summarised from previous tables as indicated. Av/11 values are averages over only those eleven categories which were shown by the learning machine approach to be linearly separable in the chosen pattern space.

to the observation that four is the number of nitrogens in a purine nucleoside, although an explanation along these lines would not be confirmed by the only middling prediction on, say, categories Pur and Adn.

Some structural classes are markedly easier to classify than others and a good guide to this is convergence of the error correction feedback procedure. The Av/11 line of table 10.1 gives the average figures of merit for the eleven categories for which the learning machine converged, and as can be seen these are for all four methods higher than those for the full twenty-one analyses. The best average value of 0.304 for the distance from the mean method comes close to Rotter and Varmuza's [145] average performance, and if the worst classified category of the eleven, C7, is deleted to leave a set of ten structural features average classification rises to 0.326. Thus if only selected categories had been presented these results could be made to appear very much better than they actually are. As might be expected linear separability of class members and non members in the training set as evidenced by convergence of the learning machine approach is a good guide to ease of classification by other methods.

It is obvious that the learning machine and k-nearest neighbour approaches achieve perfect recognition

Dimens. Preproc.	Stat. lin. discrim. 24 log	lma 24 log	Distmean ddz 0.0 61.2 log	knn k=1,ΣV 82 bin
Tr76	96.5%	99.9%	91.1%	99.9%
Pr49	74.6	75.0	79.0	61.4

Table 10.3: Overall percentage success (P_{tot}) averages for pattern recognition methods. Figures correspond to the averages of table 10.1 for the prediction set Pr49 and table 10.2 for the training set Tr76, for twenty-one analyses.

on the training set, as necessitated by the nature of their training processes, and that the method which achieves worst recognition, distance from the mean ($M_{av.}$ = 0.582 from table 10.2) is also that which performs best for prediction. Note however that this still

corresponds to an overall recognition rate of 91.1% (table 10.3). In summary, perfect recognition of those spectra on which the classifier is formed is not a prerequisite for prediction success, and in fact the reverse can often hold as demonstrated here.

The conclusion of Jurs and Isenhour with regard to such pattern recognition studies still stands that:

> "Since prediction was not perfect using any of the methods, one may infer that the information sought either does not exist or is stored in a manner which was not resolved by any of the ... approaches. Apparently, the information is stored in a non linear manner which can be approximated by a linear function, to a greater or lesser degree of accuracy, depending on the information sought"[143].

Appendix I

NUCLEOSIDE DATA BASE

The data base described in subsection 5.2.1 consists of the 70 eV electron impact mass spectra of the following 125 largely underivatised nucleosides. The systematic (CAS) name is given first, followed by common or trivial name(s), if any. If a compound has been recorded more than once, references to the spectra are in chronological order.

1. 2'-deoxy-6-thioinosine [245]
2. 1,2-dihydro-2-oxoadenosine,
   isoguanosine [284]
3. 7-(2-O-methyl-β-D-ribofuranosyl)-7H-pyrrolo[2,3-d]
   pyrimidin-4-amine,
   2'-O-methyltubercidin [241]
4. 7-(3-O-methyl-β-D-ribofuranosyl)-7H-pyrrolo[2,3-d]
   pyrimidin-4-amine,
   3'-O-methyltubercidin [241]
5. N-(3-methyl-2-butenyl)amino-3-(β-D-ribofuranosyl)-
   1H-pyrazolo[4,3-d]pyrimidin-7-amine [258,285]
6. N-(phenylmethyl)adenosine [286,285]
7. 3'-O-methyluridine [243,241]
8. 3'-O-methyladenosine [240,241]
9. 2'-deoxyadenosine [242,240,245]
10. 5'-deoxyadenosine [240]
11. 9-(β-D-psicofuranosyl)-9H-purin-6-amine,
    psicofuranine [240]
12. 9-(3-deoxy-3-fluoro-β-D-xylofuranosyl)-9H-purin-6-amine [287]
13. 9-(3-azido-3-deoxy-β-D-xylofuranosyl)-9H-purin-6-amine [287]
14. 3'-amino-3'-deoxyadenosine [288,287]
15. 9-(3-amino-3-deoxy-α-L-erythro-furanosyl)-9H-purin-6-amine [289]
16. N,N-dimethyladenosine [261,240]
17. 3'-amino-3'-deoxy-N,N-dimethyladenosine [290]
18. 3'-acetylamino-3'-deoxy-2',5'-di-O-acetyl-N,N-dimethyladenosine [290]

19. (S)-3'-[(2-amino-3-(4-methoxyphenyl)-1-oxopropyl)amino]-3'-deoxy-N,N-dimethyladenosine,
puromycin [290]
20. (S)-3'-deoxy-3'-[(2-dimethylamino-3-(4-methoxyphenyl)-1-oxopropyl)amino]-N,N-dimethyladenosine,
N,N-dimethylpuromycin [290]
21. (S)-3'-deoxy-2',5'-di-O-acetyl-3'-[(2-dimethylamino-3-(4-methoxyphenyl)-1-oxopropyl)amino]-N,N-dimethyladenosine,
2',5'-di-O-acetyl-N,N-dimethylpuromycin [290]
22. (E)-N-(4-hydroxy-3-methyl-2-butenyl)adenosine,
zeatin riboside [291,284]
23. N-(2,3-dihydroxy-3-methylbutyl)adenosine [268]
24. N-(3-methyl-3-butenyl)adenosine [284]
25. 6-chloro-9-(2-deoxy-β-D-erythro-pentofuranosyl)-9H-purine [245]
26. 5,6-dihydrouridine [261]
27. uridine [242,262,243]
28. 2'-O-methyluridine [243,241]
29. 5-methyluridine [261]
30. 2'-deoxyuridine [242]
31. 5-(hydroxymethyl)uridine [284]
32. 3-methyluridine [261]
33. 4-methoxy-1-(3-O-methyl-β-D-ribofuranosyl)-2(1H)-pyrimidinone [241]
34. 5'-deoxy-5'-(1-thyminyl)thymidine [292]
35. 2',5'-dideoxy-5'-(1-thyminyl)uridine [292]
36. 2'-O-methylguanosine [284,243,241]
37. 2'-O-methylcytidine [243,241]
38. 3'-O-methylcytidine [243,241]
39. 5-methylcytidine [261]
40. N-acetylcytidine [261]
41. adenosine [242,240,243]
42. 2'-O-methyladenosine [240,284,243,241]
43. 1,4-dihydro-3-(β-D-ribofuranosyl)-7H- pyrazolo[4,3-d]pyrimidin-7-one,
formycin B [263]
44. (S)-1-C-(7-amino-1H-pyrazolo[4,3-d]pyrimidin-3-yl)-1,4-anhydro-3-O-methyl D-ribitol,
3'-O-methylformycin [241]

45. 4-hydroxy-3-(β-D-ribofuranosyl)-1H-pyrazole-5-carboxamide, pyrazomycin [264]
46. 3-(β-D-ribofuranosyl)-1H-pyrrole-2,5-dione, showdomycin [263]
47. 3-(β-D-erythrofuranosyl)-1-(4-nitrophenyl)-1H-pyrazole [293]
48. 3-(2,3-di-O-acetyl-β-D-erythrofuranosyl)-1-(4-nitrophenyl)-1H-pyrazole [293]
49. 3-(β-D-erythrofuranosyl)-1-(4-nitrophenyl)-5-phenyl-1H-pyrazole [293]
50. N-(4-hydroxy-3-methylbutyl)adenosine [284]
51. N-(3-hydroxy-3-methylbutyl)adenosine [284]
52. 9-(β-D-allopyranosyl)-9H-purin-6-amine [240]
53. 2'-O-methyl-N-(3-methyl-2-butenyl)adenosine [285]
54. N-acetyl-3',5'-di-O-acetyl-2'-deoxy-7,8-dihydro-8-oxoadenosine [294]
55. N-(3-methyl-2-butenyl)-7-(β-D-ribofuranosyl)-7H-pyrrolo[2,3-d]pyrimidin-4-amine [258]
56. 5-(2-O-methyl-β-D-ribofuranosyl)-2,4(1H,3H)-pyrimidinedione, 2'-O-methylpseudouridine [241]
57. N-(3-methylbutyl)adenosine [295]
58. (S)-1-C-(7-amino-1H-pyrazolo[4,3-d]pyrimidin-3-yl)-1,4-anhydro-2-O-methyl-D-ribitol, 2'-O-methylformycin [241]
59. 1-(2-deoxy-β-D-erythro-pentopyranosyl)-5-methyl-2,4(1H,3H)-pyrimidinedione [296]
60. 4-methoxy-1-(2,3,5-tri-O-methyl-β-D-ribofuranosyl)-2(1H)-pyrimidinone [241]
61. 5-(β-D-ribofuranosyl)-2,4(1H,3H)-pyrimidinedione, pseudouridine [261,262]
62. (S)-1-C-(7-amino-1H-pyrazolo[4,3-d]pyrimidin-3-yl)-1,4-anhydro-D-ribitol, formycin [263]
63. 2'-deoxy-N-(phenylmethyl)adenosine [245]
64. 9-(2-deoxy-β-D-erythro-pentofuranosyl)-6-[((4-nitrophenyl)methyl)thio]-9H-purine [245]
65. 9-(2-deoxy-β-D-erythro-pentofuranosyl)-6-methylthio-9H-purine [245]
66. 2'-deoxy-N,N-dimethyladenosine [245]
67. 9-(2-deoxy-β-D-erythro-pentofuranosyl)-6-fluoro-9H-purine [245]
68. 2'-deoxy-N-hydroxyadenosine [241]

69. 6-chloro-9-(2-O-methyl-β-D-ribofuranosyl)-9H-purine [241]

70. 9-(2-O-methyl-β-D-ribofuranosyl)-6-[((4-nitrophenyl)methyl)thio]-9H-purine [241]

71. 2'-O-methyl-6-thioinosine [241]

72. 6-chloro-9-(3-O-methyl-β-D-ribofuranosyl)-9H-purine [241]

73. 9-(3-O-methyl-β-D-ribofuranosyl)-6-[((4-nitrophenyl)methyl)thio]-9H-purine [241]

74. 3'-O-methyl-6-thioinosine [241]

75. 9-[3-O-acetyl-2-chloro-2-deoxy-5-O-(2,2-dimethyl-1-oxopropyl)-β-D-arabinofuranosyl]-N-(2,2-dimethyl-1-oxopropyl)-9H-purin-6-amine [298]

76. 9-[2-O-acetyl-3-chloro-3-deoxy-5-O-(2,2-dimethyl-1-oxopropyl)-β-D-xylofuranosyl]-N-(2,2-dimethyl-1-oxopropyl)-9H-purin-6-amine [298]

77. 9-[2-chloro-2-deoxy-3-O-(4,4-dimethyl-3-(2,2-dimethyl-1-oxopropyl)oxy-1-oxo-2-pentenyl)-5-O-(2,2-dimethyl-1-oxopropyl)-β-D-arabinofuranosyl]-N-(2,2-dimethyl-1-oxopropyl)-9H-purin-6-amine [298]

78. 9-[3-chloro-3-deoxy-2-O-(4,4-dimethyl-3-(2,2-dimethyl-1-oxopropyl)oxy-1-oxo-2-pentenyl)-5-O-(2,2-dimethyl-1-oxopropyl)-β-D-xylofuranosyl]-N-(2,2-dimethyl-1-oxopropyl)-9H-purin-6-amine [298]

79. 9-[2-deoxy-3-O-(4,4-dimethyl-3-(2,2-dimethyl-1-oxopropyl)oxy-1-oxo-2-pentenyl)-5-O-(2,2-dimethyl-1-oxopropyl)-2-iodo-β-D-arabinofuranosyl]-N-(2,2-dimethyl-1-oxopropyl)-9H-purin-6-amine [298]

80. 9-[3-deoxy-2-O-(4,4-dimethyl-3-(2,2-dimethyl-1-oxopropyl)oxy-1-oxo-2-pentenyl)-5-O-(2,2-dimethyl-1-oxopropyl)-3-iodo-β-D-xylofuranosyl]-N-(2,2-dimethyl-1-oxopropyl)-9H-purin-6-amine [298]

81. 1-(2-deoxy-α-D-erythro-pentofuranosyl)-2(1H)-pyridinone [299]

82. N-(3-methyl-2-butenyl)-2-(methylthio)adenosine [300-302,284]

83. 3'-O-methyl-N-(3-methyl-2-butenyl)adenosine [285]

84. 5-(3-O-methyl-β-D-ribofuranosyl)-2,4(1H,3H)-pyrimidinedione, 3'-O-methylpseudouridine [241]

85. 6-[(3-methyl-2-butenyl)thio]-9-(β-D-ribofuranosyl)-9H-purine [285]

86. thymidine [296]

87. N-(3-methyl-2-butenyl)adenosine [303,302,284,258,295]

88. N^6-(3-methyl-2-butenyl)-9-(β-D-ribofuranosyl)-9H-purine-2,6-diamine [304]

89. 5-[[acetyl-(4,5-cis-dimethoxy-(2-cyclopenten-1-yl))amino]methyl]-N,N-dimethyl-4-methoxy-7-(2,3,5-tri-O-methyl-β-D-ribofuranosyl)-7H-pyrrolo [2,3-d]pyrimidin-2-amine [305]

90. 3'-O-methylguanosine [241]
91. 2'-deoxy-N-(3-methyl-2-butenyl)adenosine [285]
92. 3'-deoxyadenosine,
cordycepin [244,306]
93. N-(3-methylbutyl)-2-(methylthio)adenosine [295]
94. cytidine [243,267]
95. 2-methyladenosine [284]
96. 4-O-(methyloxime)uridine [307]
97. 1,9-dihydro-1-methyl-9-(2-O-methyl-β-D-ribofuranosyl)-6H-purin-6-imine [308]
98. N-methyl-2'-O-methyladenosine [308]
99. 9-(β-D-glucofuranosyl)-N-(phenylmethyl)-9H-purin-6-amine [309]
100. N-[((9-(β-D-ribofuranosyl)-9H-purin-6-yl)amino)carbonyl]glycine [310]
101. 3'-(acetylamino)-3'-deoxyadenosine [288]
102. N-acetyl-N-ethyl-2',3',5'-tri-O-ethylcytidine [311]
103. N-(4-hydroxy-3-methyl-2-butenyl)-2-(methylthio)adenosine [312,313]
104. N-methyladenosine [261]
105. 1,9-dihydro-1-methyl-9-(β-D-ribofuranosyl)-6H-purin-6-imine [261]
106. 2-methoxyadenosine [261]
107. 1,4-dihydro-5-(β-D-ribofuranosyl)-7H-pyrazolo[4,3-d] pyrimidin-
7-one [314]
108. 1,5-dihydro-6-(β-D-ribofuranosyl)-4H-pyrazolo[3,4-d]pyrimidin-4-one [314]
109. 5-(2-deoxy-β-D-erythro-pentofuranosyl)-1,4-dihydro-7H-pyrazolo[4,3-d]
pyrimidin-7-one [314]
110. 6-(2-deoxy-β-D-erythro-pentofuranosyl)-1,5-dihydro-4H-pyrazolo[3,4-d]
pyrimidin-4-one [314]
111. 4-methoxy-1-(2-O-methyl-β-D-ribofuranosyl)-2(1H)-pyrimidinone [241]
112. N^5-(2,5-anhydro-4-deoxy-D-ribohexonoyl)-4,5,6-pyrimidinetriamine [315]
113. 8-(3-deoxy-β-D-erythro-pentofuranosyl)-9H-purin-6-amine [315]
114. 8-(3-deoxy-2,5-di-O-acetyl-β-D-erythro-pentofuranosyl)-9H-purin-6-
amine [315]
115. 6-(6-amino-9H-purin-9-yl)-1,4-anhydro-6-deoxy-D-glucitol [316]
116. 1,4-anhydro-6-deoxy-6-(3,4-dihydro-2,4-dioxo-1(2H)-pyrimidinyl)-D-
glucitol [317,318]
117. 1,4-anhydro-6-deoxy-6-(3,4-dihydro-5-methyl-2,4-dioxo-1(2H)-pyrimidinyl)-
D-glucitol [317,318]
118. 6-(4-amino-2-oxo-1(2H)-pyrimidinyl)-1,4-anhydro-6-deoxy-D-glucitol
[317,318]

119. 1-C-(6-amino-9H-purin-9-yl)-2,5-anhydro-1-S-ethyl-1-thio-D-xylitol [319]

120. 1-C-(6-amino-9H-purin-9-yl)-2,5-anhydro-1-S-(2-methylpropyl)-1-thio-D-xylitol [319]

121. 2,5-anhydro-1-C-[6-(benzoylamino)-9H-purin-9-yl]-3,4-di-O-acetyl-1-S-ethyl-1-thio-D-xylitol [319]

122. 2,5-anhydro-1-C-[6-(benzoylamino)-9H-purin-9-yl]-3,4-di-O-acetyl-1-S-(2-methylpropyl)-1-thio-D-xylitol [319]

123. 1-C-(6-amino-9H-purin-9-yl)-2,5-anhydro-3,4-di-O-acetyl-1-S-(2-methylpropyl)-1-thio-D-xylitol [319]

124. N,N-dimethyl-2'-O-methylcytidine [267]

125. N-(2-furanylmethyl)adenosine,
kinetin riboside [320]

Appendix II

PROGRAMS

Programs NUCL of chapter 4 and KNNCLASSIF of chapters 8 and 9 are reproduced here together with sample outputs. The listing of program NUCL excludes the externally supplied subroutine MOLION, although the output of this subroutine has been included in the sample output. Program KNNCLASSIF generates approximately six times as much output as has been reproduced here. Only the autoscaled logarithmic variant with 82 m/z positions is included, as the other five variants, 24 and 8-14 m/z positions and the three binary analyses, are of exactly the same form.

The third listing is of program CLASSIFMEASURE (subsection 5.3.3). The computer generated tables of chapters 6-9 and the histograms of classifier performance in chapters 6-10 have been output by this program.

Program NUCL (Host)

```
 1000 BEGIN
 2000    INTEGER NOCL,NOPE,HP,MWSEARCH,SUGANAL,REVDATA ISPEC,MASSDEP,PAMI,
 3000    SPLIST,ISHRT,MININT,CL124,CL567,CL8 ;
 4000    ARRAY INT[0:1000],MASS[0:300],PR,PRMW4,TH,MW4[0:21] ;
 5000    ALPHA ARRAY CMMNT,NAME,CMMNTSOT[0:12] ;
 6000    FILE FILE6 (KIND=PRINTER) ; DEFINE LP=FILE6# ;


 7000    PROCEDURE HPCLSTR(X) ; PROCEDURE X ; EXTERNAL ;

 8000    PROCEDURE ORDMW(M,P,N) ; ARRAY M,P[*] ; INTEGER N ; EXTERNAL ;

 9000    PROCEDURE MWLOSS(M,P,N,I,PM,SM) ;
10000    ARRAY M,P,PM,SM[*] ; INTEGER N,I ; EXTERNAL ;

11000    PROCEDURE MOLION(A,B,C,D) ; ARRAY A,B[*] ; INTEGER C,D ; EXTERNAL ;

12000    PROCEDURE CLASSGT ; EXTERNAL ;

13000    PROCEDURE CLASSGTB(A,B) ; ARRAY A,B[*] ; EXTERNAL ;

14000    PROCEDURE CLSPEC1(A,B) ; ARRAY A,B[*] ; EXTERNAL ;

15000    PROCEDURE CLSPEC2(A B) ; ARRAY A,B[*] ; EXTERNAL ;

16000    PROCEDURE CLSPEC4 ; EXTERNAL ;

17000    PROCEDURE CLSPEC5(A,B) ; ARRAY A,B[*] ; EXTERNAL ;

18000    PROCEDURE CLASSGTC(A,B) ; ARRAY A,B[*] ; EXTERNAL ;

19000    PROCEDURE FILAR(A,B) ; REAL A ; POINTER B ; EXTERNAL ;

20000    PROCEDURE FILARVAL(A,B) ; REAL A ; POINTER B ; EXTERNAL ;
21000    DEFINE PRFACT(A)= 0.001*(A)*INT[(A)] # ;
22000
23000    REAL I,J,K,DTFRM,NOSP,ITAPE,NSOT ,W,COUNT,OLD,BIGST,INDEX,OUTCHARS,
24000    TPRNT,INCHARS,NOTP ;
25000    ARRAY SOT[0:15] ;
26000    FILE CR(KIND=READER) ;

27000    PROCEDURE EXEC ; EXTERNAL ;

28000    PROCEDURE ZERO ;
29000    BEGIN
30000       OLD:=BIGST:=NOPE:=HP:=COUNT:=J:=0 ;
31000       DO VECTORMODE (PR,MW4,PRMW4,FOR 22)
32000       BEGIN PR:=MW4:=PRMW4:=0 ; INCREMENT PR,MW4,PRMW4 ; END ;
33000       DO VECTORMODE (INT,FOR 1001) BEGIN INT:=0 ; INCREMENT INT ; END ;
34000       DO VECTORMODE (MASS,FOR 301) BEGIN MASS:=0 ; INCREMENT MASS ; END ;
35000    END ;
36000
37000    %         ASSORTED DATA
38000    READ (CR, <3I5>, NOTP,NOSP,NOCL ) ;                                  % READ
39000    READ (CR, <16I5>, FOR I:=1 STEP 1 UNTIL NOCL DO TH[I] ) ;             % READ
40000
41000    WRITE (LP, <"NO.OF PAPERTAPES =",I4,/>, NOTP ) ;
42000    WRITE (LP, <"NO.OF SPECTRA FROM CARDS =",I4,/>, NOSP ) ;
43000    DO VECTORMODE (CMMNTSOT,FOR 13)
44000    BEGIN CMMNTSOT:="       " ; INCREMENT CMMNTSOT ; END ;
45000
46000    %         TAPE INPUT
47000
48000    IF NOTP >0 THEN
49000    BEGIN
50000       BOOLEAN ARRAY B[0:10000] ;
51000       POINTER POUT,PIN,PA,PB ;
52000       LABEL L2 ;
53000       BOOLEAN ASCIIZONE,GOTIT ;
54000       BOOLEAN ARRAY BUF[0:13],OUTBUF[0:29] ;
55000       TRUTHSET NUM("0123456789.$ "") ;

56000       PROCEDURE ASN(U,V) ;
57000       REAL U,V ;
58000       IF PA-U EQL "1" THEN W:=*+1.0*10.0**V ELSE
59000       IF PA-U EQL "2" THEN W:=*+2.0*10.0**V ELSE
60000       IF PA-U EQL "3" THEN W:=*+3.0*10.0**V ELSE
61000       IF PA-U EQL "4" THEN W:=*+4.0*10.0**V ELSE
62000       IF PA-U EQL "5" THEN W:=*+5.0*10.0**V ELSE
63000       IF PA-U EQL "6" THEN W:=*+6.0*10.0**V ELSE
64000       IF PA-U EQL "7" THEN W:=*+7.0*10.0**V ELSE
65000       IF PA-U EQL "8" THEN W:=*+8.0*10.0**V ELSE
66000       IF PA-U EQL "9" THEN W:=*+9.0*10.0**V ;
67000
```

```
 68000        FOR ITAPE:=1 STEP 1 UNTIL NOTP DO
 69000        BEGIN
 70000           INDEX:=OUTCHARS:=INCHARS:=0 ;
 71000           FOR I:=0,1,2,3,4,5,6,7,8,9,10,11,12,13 DO BUF[I]:=FALSE ;
 72000           FOR I:=0 STEP 1 UNTIL 29 DO OUTBUF[I]:=FALSE ;
 73000           DO VECTORMODE (SOT,FOR 11) BEGIN SOT:=0 ; INCREMENT SOT ; END ;
 74000           FOR I:=0 STEP 1 UNTIL 10000 DO B[I] := FALSE ;
 75000
 76000           READ (CR, <13A6>, CMMNT ) ;                                    % READ
 77000           READ (CR,<11I5>, MWSEARCH,SUGANAL,TPRNT,MASSDEF,PAMI,SPLIST,   % READ
 78000           ISHRT, MININT, CL124, CL567, CL8 ) ;
 79000           READ (CR, <16I5>, NSOT,FOR I:=1 STEP 1 UNTIL NSOT DO SOT[I] ) ; % READ
 80000           READ (CR, <13A6>, NAME ) ;                                     % READ
 81000
 82000           PA := POINTER(NAME) ;
 83000           SCAN PA:PA FOR 78 UNTIL EQL " " ;
 84000           REPLACE PA BY "." ;
 85000           BEGIN
 86000              FILE PT3(KIND=1,TITLE=NAME,FILETYPE=7) ;
 87000              K := 0 ;
 88000              OUTCHARS := 180 ;
 89000              ASCIIZONE := BOOLEAN(4"7F7F7F7F7F7F") ;
 90000              REPLACE POUT := POINTER(OUTBUF) BY " " FOR 30 WORDS ;
 91000              WHILE NOT READ (PT3,80,BUF) DO
 92000              BEGIN
 93000                 FOR INDEX := 0 STEP 1 UNTIL 13 DO BUF[INDEX] := * AND ASCIIZONE ;
 94000                 REPLACE POUT:POUT BY PIN:POINTER(BUF) FOR (INCHARS := MIN(80,OUTCHARS) )
 95000                 WITH ASCIITOEBCDIC ;
 96000                 IF (OUTCHARS := *-INCHARS) LEQ 0 THEN
 97000                 BEGIN
 98000                    FOR J:=K STEP 1 UNTIL K+29 DO B[J] := OUTBUF[J-K] ; K := K+30 ;
 99000                    REPLACE POUT := POINTER(OUTBUF) BY " " FOR 30 WORDS ;
100000                    REPLACE POUT : POUT BY PIN : PIN FOR (80-INCHARS) WITH ASCIITOEBCDIC ;
101000                    OUTCHARS := 100+INCHARS ;
102000                 END ;
103000              END ;
104000              IF OUTCHARS LSS 180 THEN
105000              BEGIN
106000                 FOR J:=K STEP 1 UNTIL K+29 DO B[J] := OUTBUF[J-K] ; K := K+30 ;
107000              END ;
108000              INDEX := -1 ;
109000              WHILE (GOTIT := REAL (OUTBUF[INDEX := *+1] ) ISNT 48"070707070707" )
110000              AND INDEX LSS 29 DO ;
111000              IF GOTIT THEN
112000              BEGIN
113000                 REPLACE OUTBUF[INDEX] BY " " FOR 30-INDEX WORDS ;
114000                 REPLACE B[K] BY " " FOR 30 WORDS ;
115000                 K:=*+30 ;
116000              END ;
117000              WRITE (LP[SKIP 1] ) ;
118000              WRITE(LP,<13A6,//>,FOR I:=0 STEP 1 UNTIL 12 DO CMMNT[I]) ;
119000              WRITE (LP, <13A6,///>, FOR I:=0 STEP 1 UNTIL 12 DO NAME[I] ) ;
120000              IF TPRNT=1 THEN
121000              BEGIN
122000                 FOR J:=0 STEP 10 UNTIL K+10 DO
123000                 WRITE (LP, <I4,X2,10A6>, J,FOR I:=J STEP 1 UNTIL J+9 DO B[I] ) ;
124000                 WRITE (LP,</>) ;
125000                 FOR J:=0 STEP 10 UNTIL K+10 DO
126000                 WRITE (LP, <I4,X2,10H12>,J,FOR I:=J STEP 1 UNTIL J+9 DO B[I] ) ;
127000                 WRITE (LP,</>) ;
128000              END ;
129000
130000              %        CHECK RAW DATA + REMOVE RUBBISH
131000              PA:=PB:=POINTER(B) ;
132000              I := 0 ;
133000              WHILE I<6*K+60 DO
134000              BEGIN
135000                 IF PA IN NUM THEN REPLACE PB:PB BY PA:PA FOR 1
136000                 ELSE
137000                 PA := *+1 ;
138000                 I := *+1 ;
139000              END ;
140000              REPLACE PB BY "$" ;
141000              %        CHECK >=4 CHAR BEFORE FIRST "."
142000              SCAN POINTER(B) FOR J:4 UNTIL EQL "." ;
143000              IF J>0 THEN
144000              BEGIN
145000                 PA := POINTER(B[K+10]) ;
146000                 THRU 6*K+60 DO BEGIN REPLACE PA+J BY PA FOR 1 ; PA:=*-1 ; END ;
147000                 REPLACE PA+J BY PA FOR 1 ;
148000                 PA:=POINTER(B) ;
149000                 REPLACE PA BY " " FOR J ;
150000              END ;
```

```
151000            IF TPRNT=1 THEN
152000            BEGIN
153000               FOR J:=0 STEP 10 UNTIL K+10 DO
154000               WRITE (LP, <I4,X2,10A6>, J,FOR I:=J STEP 1 UNTIL J+9 DO B[I] ) ;
155000               WRITE (LP, <///> ) ;
156000            END ;
157000         END ;
158000         %        EXTRACT MASS,INT VALUES
159000         ISPEC := 0 ;
160000         PA := POINTER(B) ;
161000         L2 : ZERO ;
162000         ISPEC := *+1 ;
163000         WHILE PA NEQ "$" AND PA NEQ 48"00" AND J<300 DO
164000         IF PA EQL "." THEN
165000         BEGIN
166000            J := *+1 ;
167000            FOR I:=1,2,3,4 DO ASN(I,I-1) ;
168000            PA := *+6 ;
169000            FOR I:=1,2,3,4,5 DO ASN(I,I-6) ;
170000            IF J<300 THEN MASS[J] := W
171000            ELSE
172000            BEGIN WRITE(LP, <">300 MASS ON TAPE",I3,F10.2,///>,J,W) ; J:=300 ; END ;
173000            W := 0 ;
174000            PA := *+4 ;
175000            FOR I:=1,2,3,4 DO ASN(I,I-1) ;
176000            IF MASS[J]<1000 THEN INT[J] := W
177000            ELSE
178000            WRITE (LP, <"TAPE ERROR MASS >1000",I3,2F10.2,///>, J,MASS[J],W) ;
179000            W := 0 ;
180000         END
181000         ELSE
182000         PA := *+1 ;
183000         NOPE := J ;
184000
185000         IF NSOT=0 THEN EXEC
186000         ELSE
187000         FOR I:=1 STEP 1 UNTIL NSOT DO
188000         IF SOT[I]=ISPEC THEN
189000         BEGIN
190000            READ (CR, <13A6>, CMMNTSOT ) ;                                    % READ
191000            EXEC ;
192000            DO VECTORMODE (CMMNTSOT,FOR 13)
193000            BEGIN CMMNTSOT:="      " ; INCREMENT CMMNTSOT ; END ;
194000         END ;
195000
196000         IF PA:=*+1 IN NUM THEN GO TO L2 ;
197000      END ;
198000   END ;
199000   MASSDEF := 0 ;
200000   DO VECTORMODE (NAME,FOR 13) BEGIN NAME:="      " ; INCREMENT NAME ; END;
201000
202000   %        CARD INPUT
203000
204000   IF NOSP>0 THEN
205000   READ (CR, <16I5>, MWSEARCH, SUGANAL,REVDATA, PAMI, SPLIST,          % READ
206000   ISHRT, MININT, CL124, CL567, CL8 ) ;
207000   FOR ISPEC:=1 STEP 1 UNTIL NOSP DO
208000   BEGIN
209000      ZERO ;
210000
211000      READ (CR, <13A6>, CMMNT ) ;                                        % READ
212000      READ (CR, <2I5>, NOPE, DTFRM                                       % READ
213000
214000      %        CARD INPUT FORMATS
215000      CASE DTFRM OF
216000      BEGIN
217000         READ (CR, <20I4>, FOR I:=1 STEP 1 UNTIL NOPE DO                    % READ
218000         [ MASS[I], INT[I] ] ) ;
219000         BEGIN
220000            READ (CR, <16I5>, FOR I:=1 STEP 1 UNTIL NOPE DO MASS[I] ) ;     % READ
221000            FOR I:=1 STEP 1 UNTIL NOPE DO INT[I] := 100 ;
222000         END ;
223000         READ (CR, <16I5>, FOR I:=1 STEP 1 UNTIL NOPE DO                    % READ
224000         [ MASS[I], INT[I] ] ) ;
225000         READ (CR,<8F10.0>,FOR I:=1 STEP 1 UNTIL NOPE DO                    % READ
226000         [ MASS[I], INT[I] ] ) ;
227000         READ (CR,/, FOR I:=1 STEP 1 UNTIL NOPE DO [ MASS[I],INT[I] ] ) ;  % READ
228000      END ;
229000
230000      EXEC ;
231000
232000   END ;
233000 END .
```

Program NUCL (Procedures)

```
 1000 [
 2000 INTEGER NOCL,NOPE,HP,MWSEARCH,SUGANAL,REVDATA,ISPEC,MASSDEF,PAMI,
 3000 SPLIST,ISHRT,MININT,CL124,CL567,CL8 ;
 4000 ARRAY INT[0:1000],MASS[0:300],PR,PRMW4,TH,MW4[0:21] ;
 5000 ALPHA ARRAY CMMNT,NAME,CMMNTSOT[0:12] ;
 6000 FILE FILE6 (KIND=PRINTER) ; DEFINE LP=FILE6# ;


 7000 PROCEDURE HPCLSTR(X) ; PROCEDURE X ; EXTERNAL ;

 8000 PROCEDURE ORDMW(M,P,N) ; ARRAY M,P[*] ; INTEGER N ; EXTERNAL ;

 9000 PROCEDURE MWLOSS(M,P,N,I,PM,SM) ;
10000 ARRAY M P,PM,SM[*] ; INTEGER N,I ; EXTERNAL ;

11000 PROCEDURE MOLION(A,B,C,D) ; ARRAY A,B[*] ; INTEGER C,D ; EXTERNAL ;

12000 PROCEDURE CLASSGT ; EXTERNAL ;

13000 PROCEDURE CLASSGTB(A,B) ; ARRAY A,B[*] ; EXTERNAL ;

14000 PROCEDURE CLSPEC1(A,B) ; ARRAY A,B[*] ; EXTERNAL ;

15000 PROCEDURE CLSPEC2(A,B) ; ARRAY A,B[*] ; EXTERNAL ;

16000 PROCEDURE CLSPEC4 ; EXTERNAL ;

17000 PROCEDURE CLSPEC5(A,B) ; ARRAY A,B[*] ; EXTERNAL ;

18000 PROCEDURE CLASSGTC(A,B) ; ARRAY A,B[*] ; EXTERNAL ;

19000 PROCEDURE FILAR(A,B) ; REAL A ; POINTER B ; EXTERNAL ;

20000 PROCEDURE FILARVAL(A,B) ; REAL A ; POINTER B ; EXTERNAL ;
21000 DEFINE PRFACT(A)= 0.001*(A)*INT[(A)] # ;
22000 ]
23000


24000 PROCEDURE CLASSGTC(CGLYC,CGLYCPR) ;
25000
26000 %        CARBON GLYCOSIDE ASSIGNMENT
27000 %        DETERMINE CLASS PR AND BASE WT ONLY
28000 %        MW ESTIMATE IN CLASSGTB
29000 %        NO CORRESPONDING CLASS SPECIFIC PROC
30000
31000 ARRAY CGLYC,CGLYCPR[*] ;
32000 BEGIN
33000    REAL I,J,B,ICG,FLG,MS,BP ;
34000    ALPHA ARRAY OPUTLINE[0:20] ;
35000    POINTER PA ;
36000    FORMAT LYNE(130("**"),/) ;
37000    WRITE (LP, LYNE) ;
38000    ICG := 0 ;
39000    FOR I:=1 STEP 1 UNTIL NOPE DO
40000    IF MASS[I]>=110 THEN
41000    BEGIN
42000       B := MASS[I]-30 ;
43000       IF INT[B+30]>=20 AND INT[B+2]<=30 AND INT[B+1]<=30 THEN
44000       BEGIN
45000          FLG := 0 ;
46000          FOR MS:=14,15,28,44,58,60,74 DO IF INT[B+MS]>0 THEN FLG:=1 ;
47000          IF FLG=1 THEN
48000          BEGIN
49000             DO VECTORMODE ( OPUTLINE, FOR 21)
50000             BEGIN OPUTLINE:="       " ; INCREMENT OPUTLINE ; END ;
51000             PA := POINTER(OPUTLINE) +1 ;
52000             FILAR(B,PA) ;
53000             REPLACE PA:PA BY " I     I " ;
54000             CGLYC[ICG:=*+1] := B ;
55000             IF ICG=1 THEN
56000             WRITE ( LP, < "CHARACTERISTIC C GLYCOSIDE IONS :", //,
57000             "BASE I  PR I    B+14      B+15      B+28      B+30      B+44      B+58
58000             B+60      B+74", /,
59000             "CAND I     I", X34, "=====", /,
60000             2 ( X5,"I" ) > ) ;
61000             FOR MS:=14,15,28,30,44,58,60,74 DO
62000             IF INT[BP:=B+MS]>0 THEN
63000             BEGIN CGLYCPR[ICG]:=*+PRFACT(BP) ; FILARVAL(BP,PA) ; END
64000             ELSE
65000             REPLACE PA:PA BY "          " ;
66000             PA := POINTER(OPUTLINE) +7 ;
67000             FILAR(CGLYCPR[ICG],PA) ;
68000             WRITE (LP, <21A6,/, 2 ( X5,"I" ) >, OPUTLINE ) ;
```

```
69000          END ;
70000       END ;
71000    END ;
72000    WRITE (LP, LYNE) ;
73000 END ;
74000

75000 PROCEDURE CLSPEC5(BWT,BPR) ;
76000
77000 %         CLASS SPECIFIC PROCEDURE FOR NUCL CLASSES 5;6,7
78000 %         AIM TO CLASSIFY BASE PART AS URD,CYT,GUN,ADN
79000 %         ACCORDING TO RECOGNISED FRAGMENTATIONS OF BASE
80000 %         TO ACCOMODATE SUBST BASES LOOK FOR BOTH
81000 %         STANDARD LOSSES FROM B+1 , AND FOR
82000 %         STANDARD IONS FORMED BY SUCH LOSSES FROM THE UNSUBST BASE
83000 %         ALSO FOR CHAR ION AT B+41 IN CASE OF CYT
84000
85000 %         TREAT THIS SECTION WITH CAUTION
86000
87000 ARRAY BWT,BPR[*] ;
88000 BEGIN
89000    FORMAT LYNE (130 "**"),/) ;
90000    ARRAY LSS[1:10] ;
91000    ALPHA ARRAY TITLE[1:3,0:14] ;
92000

93000    PROCEDURE BASETYPE(B,L,T) ;
94000
95000    %         FOR EACH BWT LOOK FOR LOSSES L
96000    %         AND CHAR IONS ARISING FROM SUCH LOSSES FROM UNSUBST BASE B
97000    %         OUTPUT UNDER TITLE T
98000
99000    REAL B ;
100000   ARRAY L[*] ;
101000   ALPHA ARRAY T[*,*] ;
102000   BEGIN
103000      REAL I,NOL,CI,NPR,IB ;
104000      ALPHA ARRAY OPUTLINE[0:21] ;
105000      POINTER PA ;
106000      DEFINE FORI= FOR I:=1 STEP 1 UNTIL NOL DO # ;
107000
108000      FOR I:=1 STEP 1 WHILE L[I]^=0 DO NOL:=*+1 ;
109000      WRITE (LP, < "CHARACTERISTIC IONS  ",
110000      * (I3,"," ), " FROM B+1 = ", I3, ") : ",
111000      6  " ",I3,",",I3,") " ), /, X67, 6  " ",I3,",",I3,") " ) >,
112000      NOL. FORI B+1-L[I], B+1,
113000      FORI IF INT[ CI:=B+1-L[I] ]>0 THEN [ CI,INT[CI] ] ) ;
114000      WRITE (LP) ;
115000      WRITE (LP, < 3(15A6,/), "     I      I      I">,
116000      FOR I:=1,2,3 DO T[I,*] ) ;
117000
118000      FOR IB:=1 STEP 1 WHILE BWT[IB]>0 DO
119000      BEGIN
120000         DO VECTORMODE (OPUTLINE,FOR 22)
121000         BEGIN OPUTLINE := "      " ; INCREMENT OPUTLINE ; END ;
122000         NPR := 0 ;
123000         PA := POINTER(OPUTLINE)+1 ;
124000         FILAR(BWT[IB],PA) ;
125000         REPLACE PA:PA BY " I  " ;
126000         FILAR(BPR[IB],PA) ;
127000         REPLACE PA:PA BY " I  " ;
128000         PA := *+3 ;
129000         REPLACE PA:PA BY " I " ;
130000         FORI IF INT[CI:=BWT[IB]+1-L[I]]>0 THEN
131000         BEGIN FILARVAL(CI,PA) ; NPR:= *+PRFACT(CI) ; END ELSE PA:=*+10 ;
132000         PA := POINTER(OPUTLINE)+15 ;
133000         FILAR(BPR[IB]+NPR,PA) ;
134000         IF NPR>0 THEN
135000         WRITE (LP, <22A6, /, "     I      I      I">, OPUTLINE ) ;
136000      END ;
137000   END ;
```

```
138000     PROCEDURE ZTL ;
139000     %         ZERO TITLES AND RULE OFF
140000     BEGIN
141000        DO VECTORMODE (TITLE[1,*], FOR 15)
142000        BEGIN TITLE:="       " ; INCREMENT TITLE ; END ;
143000        DO VECTORMODE (TITLE[2,*], FOR 15)
144000        BEGIN TITLE:="       " ; INCREMENT TITLE ; END ;
145000        DO VECTORMODE (TITLE[3,*], FOR 15)
146000        BEGIN TITLE:="       " ; INCREMENT TITLE ; END ;
147000        FILL TITLE[3,*] WITH "      I       I       I     " ;
148000        WRITE (LP, LYNE) ;
149000     END ;
150000
151000
152000
153000
154000
155000
156000
157000     %         URACIL TYPE BASE
158000
159000     ZTL ;
160000     WRITE (LP, <   "URACIL TYPE BASE ASSUMPTION :",/> ) ;
161000     FILL TITLE[1,*] WITH "BASE I ORIG I  NEW I  B+1-43   " ;
162000     FILL TITLE[2,*] WITH "CAND I   PR I   PR I  (HNCO)   " ;
163000     FILL LSS WITH 43,9(0) ;
164000     BASETYPE(111,LSS,TITLE) ;
165000
166000     %         CYTOSINE TYPE BASE
167000
168000     ZTL ;
169000     WRITE (LP, <   "CYTOSINE TYPE BASE ASSUMPTION :",/> ) ;
170000     FILL TITLE[1,*] WITH "BASE I ORIG I  NEW I  B+1-16     B+1-28     B+1-42
171000     B+41    " ;
172000     FILL TITLE[2,*] WITH "CAND I   PR I   PR I   (NH2)      (CO)       (NCO)
173000     (B-CO-CH)" ;
174000     FILL LSS WITH 16,28,42,-40,6(0) ;
175000     BASETYPE(110,LSS,TITLE) ;
176000
177000     %         GUANINE TYPE BASE
178000
179000     ZTL ;
180000     WRITE (LP, <   "GUANINE TYPE BASE ASSUMPTION : ",/> ) ;
181000     FILL TITLE[1,*] WITH "BASE I ORIG I  NEW I  B+1-16     B+1-17     B+1-41
182000     B+1-42     B+1-44     B+1-70     B+1-72   " ;
183000     FILL TITLE[2,*] WITH "CAND I   PR I   PR I   (NH2)      (NH3)     (HNCN)
184000     (NCO)    (NH2+CO)    NH2+HCN     NH2+CO " ;
185000     FILL TITLE[3,*] WITH "      I       I       I
186000     +HCN)       +CO)   " ;
187000     FILL LSS WITH 16,17,41,42,44,70,72,3(0) ;
188000     BASETYPE(150,LSS,TITLE) ;
189000
190000     %         ADENINE TYPE BASE
191000
192000     ZTL ;
193000     WRITE (LP, <   "ADENINE TYPE BASE ASSUMPTION : ",/> ) ;
194000     FILL TITLE[1,*] WITH "BASE I ORIG I  NEW I  B+1-15     B+1-16     B+1-27
195000     B+1-54   " ;
196000     FILL TITLE[2,*] WITH "CAND I   PR I   PR I    (NH)      (NH2)      (HCN)
197000     (HCN+HCN)" ;
198000     FILL LSS WITH 15,16,27,54,6(0) ;
199000     BASETYPE(134,LSS,TITLE) ;
200000
201000     WRITE (LP,LYNE) ;
202000  END ;
203000

204000  PROCEDURE FILAR(W,PA) ;
205000
206000  %         FILL ARRAY OUTPUTLINE WITH AN INTEGER NUMBER
207000
208000  REAL W ;
209000  POINTER PA ;
210000  BEGIN
211000     REAL A,TEMP,B ;
```

```
212000     PROCEDURE FIL ;
213000     CASE A OF
214000     BEGIN
215000        REPLACE PA:PA BY "0" ;
216000        REPLACE PA:PA BY "1" ;
217000        REPLACE PA:PA BY "2" ;
218000        REPLACE PA:PA BY "3" ;
219000        REPLACE PA:PA BY "4" ;
220000        REPLACE PA:PA BY "5" ;
221000        REPLACE PA:PA BY "6" ;
222000        REPLACE PA:PA BY "7" ;
223000        REPLACE PA:PA BY "8" ;
224000        REPLACE PA:PA BY "9" ;
225000     END ;
226000     B := INTEGER(W) ;
227000     IF TEMP:=A:=B DIV 100 = 0 THEN REPLACE PA:PA BY " " ELSE FIL ;
228000     IF A:=(B DIV 10) MOD 10 = 0 AND TEMP=0 THEN REPLACE PA:PA BY " " ELSE
229000     FIL ;
230000     A := B MOD 10 ;
231000     FIL ;
232000 END ;
233000

234000 PROCEDURE FILARVAL(M,PA) ;
235000
236000 %         FILL ARRAY OUTPUTLINE WITH A MASS,INT PAIR
237000 %        OF FORM "(123,456) "
238000
239000 REAL M ;
240000 POINTER PA ;
241000 BEGIN
242000     REPLACE PA:PA BY "(" ;
243000     FILAR(M,PA) ;
244000     REPLACE PA:PA BY "," ;
245000     FILAR(INT[M],PA) ;
246000     REPLACE PA:PA BY ") " ;
247000 END ;
248000

249000 PROCEDURE CLASSGTB(BCND,BCNDPR) ;
250000
251000 %         DETERMINE TYPE OF SUGAR AND MW OF NUCLEOSIDES
252000
253000 ARRAY BCND,BCNDPR[*] ;
254000 BEGIN
255000     FORMAT LYNE  130("**"),/) ;
256000     REAL I,J,IMW,V,W,MW,PR,CNT,MNP,FLG,IBS ;
257000     ARRAY MCND,MCNDPR[0:30], INS[0:12] ;
258000     ARRAY M133,M133PR,M147,M147PR,M117,M117PR[0:30] ;
259000     ARRAY BRIB,BRIBPR,B2OM,B2OMPR,B2DO,B2DOPR[0:30] ;
260000     POINTER PA ;
261000     ALPHA ARRAY OPUTLINE[0:21] ;
262000

263000     PROCEDURE PLACECND (M,P,MAR,PAR ) ;
264000     REAL M,P ;
265000     ARRAY MAR,PAR[*] ;
266000     BEGIN
267000        REAL I,J,N ;
268000        LABEL L1 ;
269000        FOR I:=1 STEP 1 WHILE MAR[I]>0 DO N:= *+1 ;
270000        FOR I:=1 STEP 1 WHILE MAR[I]>0 DO IF M>MAR[I] THEN GO TO L1 ;
271000        L1 : N := *+1 ;
272000        FOR J:=N STEP -1 UNTIL I DO
273000        BEGIN MAR[J]:=MAR[J-1] ; PAR[J] := PAR[J-1] ; END ;
274000        MAR[I] := M ;
275000        PAR[I] := P ;
276000        MNP := MIN(MNP,P) ;
277000     END ;
278000
```

```
279000    PROCEDURE OPUT(IONS) ;
280000    %         IONS[1]=M OR B CAND , IONS[0]=CAND PR , IONS[2->]=FRAGMENT IONS
281000    ARRAY IONS[*] ;
282000    BEGIN
283000       REAL I,MS ;
284000       DO VECTORMODE (OPUTLINE, FOR 22)
285000       BEGIN OPUTLINE:="       " ; INCREMENT OPUTLINE ; END ;
286000       PA := POINTER(OPUTLINE[0])+1 ;
287000       FILAR(IONS[1],PA) ;
288000       REPLACE PA:PA BY " I " ;
289000       FILAR(IONS[0],PA) ;
290000       REPLACE PA:PA BY " I " ;
291000       FOR I:=1 STEP 1 WHILE MS:=IONS[I]>0 DO
292000       IF INT[MS]>0 THEN FILARVAL(MS,PA)
293000       ELSE
294000       REPLACE PA:PA BY "           " ;
295000       WRITE (LP, <22A6, /, "      I      I">, OPUTLINE ) ;
296000    END ;
297000

298000    PROCEDURE RIBOSE(B30,B44,B60) ;
299000
300000    %         PICK IONS >10% REL INT AS POSSIBLE B+1 OR B+2
301000    %         AND PASS TO BASFIND
302000
303000    REAL B30,B44,B60 ;
304000    BEGIN
305000       REAL PREC ;

306000       PROCEDURE BASFIND(A12,A30,A44,A60) ;
307000
308000       %         DETERMINE B VIA B,B+1,2,30,44,60
309000       %         ALLOWING FOR 2' OME OR DEOXY
310000
311000       REAL A12,A30,A44,A60 ;
312000       BEGIN
313000          REAL CNT,MS,PR,B,GT20 ;
314000          IF D:=MASS[I]-A12 >= 80 THEN
315000          BEGIN
316000             GT20 := CNT := 0 ;
317000             %         LOOK FOR IONS AT B,B+1,B+2,B+30,B+44(58 OR 28),B+60(74 OR -)
318000             FOR MS:=B,B+1,B+2,B+A30,B+A44,IF A60=0 THEN 0 ELSE B+A60 DO
319000             IF INT[MS]>0 THEN
320000             BEGIN CNT:=*+1 ; IF INT[MS] >= 20 THEN GT20 := 1 ; END ;
321000             %         CRITERIA : >=3 AND ONE >=20%, OR ,B+1 >=40%
322000             IF (CNT>=3 AND GT20=1) OR INT[B+1]>=40 THEN
323000             BEGIN
324000                FOR MS:=B,B+1,B+2,B+14,B+15,B+28,B+30,B+44,B+58,B+60,B+74 DO
325000                PR := *+PREFACT(MS) ;
326000                IF IBS<60 THEN BEGIN BCND[IBS:=*+1]:=B ; BCNDPR[IBS]:=PR ; END
327000                ELSE
328000                WRITE (LP, < ">60 B =",2I5>, B,PR) ;
329000             END ;
330000          END ;
331000       END ;
332000
333000       FILL BCND   WITH 31(0) ;
334000       FILL BCNDPR WITH 31(0) ;
335000       IBS := 0 ;
336000       PREC := 0 ;
337000       FOR I:=1 STEP 1 UNTIL NOPE DO
338000       IF INT[MASS[I]]>=10 THEN
339000       BEGIN
340000          IF PREC^=MASS[I]-1 THEN BASFIND(1,B30,B44,B60) ;
341000          BASFIND(2,B30,B44,B60) ;
342000          PREC := MASS[I]-2 ;
343000       END ;
```

```
344000
345000      %         OUTPUT LINE OF BASE IONS
346000      WRITE (LP, <  "CHARACTERISTIC IONS FOR B :", //,
347000      "BASE I  PR I     B         B+1        B+2        B+14       B+15       B+28
348000      B+30      B+44       B+58       B+60       B+74", /,
349000      "CAND I     I",
350000      X51, "(B+44-16)", X21, "(B+44+14)", X11, "(B+60+14)",
351000      2( /,"     I     I" ) > ) ;
352000      FOR I:=1 STEP 1 UNTIL IBS DO
353000      BEGIN
354000         J := -1 ;
355000         FOR V:=0,0,1,2,14,15,28,30,44,58,60,74 DO INS[J:=*+1] := BCND[I]+V ;
356000         INS[0] := BCNDPR[I] ;
357000         OPUT(INS) ;
358000      END ;
359000      WRITE (LP,LYNE) ;
360000   END ;
361000

362000   PROCEDURE LOSMW(SGMS,M133,M133PR) ;
363000
364000   %         FOR THE 3 STANDARD SUGAR TYPES ASSIGN MW ON BASIS OF
365000   %         LOSSES FROM M
366000
367000   ARRAY M133,M133PR[*] ;
368000   REAL SGMS ;
369000   BEGIN
370000      REAL IMW,CNT,I,PR,MW,V,MS ;
371000      WRITE (LP, <
372000      "  MW I FROM I  PR", " I      M         M-15       M-17       M-18       M-30
373000      M-31       M-32       M-35       M-36       M-61", /,
374000      "CAND I    B I     ", " I              (CH3)      (OH)      (H2O)     (H2C=O
375000      )    (CH2OH)    (CH3OH)    (OH+H2O)  (H2O+H2O)   (30+31)",
376000      2( /, "     I       I      I" ) >, SGMS ) ;
377000      IMW := 0 ;
378000      FOR I:=1 STEP 1 UNTIL IBS DO
379000      BEGIN
380000         PR := CNT := 0 ;
381000         MW := SGMS+BCND[I] ;
382000         IF MW>=MASS[3] THEN
383000         BEGIN
384000            %         LOOK FOR LOSSES FROM POSS MW
385000            FOR V:=0,15,17,18,30,31,32,35,36,61 DO
386000            IF MS:=MW-V>0 THEN IF INT[MS]>0 THEN
387000            BEGIN PR := *+PRFACT(MS) ; CNT := *+1 ; END ;
388000            %         CRITERION : >=1 SUCH LOSS
389000            IF CNT>=1 THEN
390000            IF IMW<30 THEN
391000            BEGIN
392000               M133[IMW:=*+1] := MW ;
393000               M133PR[IMW] := PR ;
394000               %         OUTPUT LINE OF LOSSES FROM CAND
395000               DO VECTORMODE (OPUTLINE, FOR 22)
396000               BEGIN OPUTLINE:="       " ; INCREMENT OPUTLINE ; END ;
397000               PA := POINTER(OPUTLINE[0]) +1 ;
398000               FILAR(MW,PA) ;
399000               REPLACE PA:PA BY " I  " ;
400000               FILAR(BCND[I],PA) ;
401000               REPLACE PA:PA BY " I " ;
402000               FILAR(PR,PA) ;
403000               REPLACE PA:PA BY " I " ;
404000               FOR V:=0,15,17,18,30,31,32,35,36,61 DO
405000               IF MS:=MW-V>0 THEN IF INT[MS]>0 THEN FILARVAL(MS,PA)
406000               ELSE
407000               REPLACE PA:PA BY "          " ;
408000               WRITE (LP, <22A6, /, "     I      I     I">, OPUTLINE) ;
409000            END
410000            ELSE
411000            WRITE (LP, < ">30 SUG=133 NUCL MW =",I4,/>, MW) ;
412000         END ;
413000      END ;
414000      WRITE (LP,LYNE) ;
415000   END ;
416000
```

```
417000    PROCEDURE BCOLLECT(B1,B1P,B2,B2P) ;
418000
419000    %        GET ONE LIST , IN B2 , OF ALL BASE CANDIDATES
420000    %        PUTTING NEW ONES FROM B1 INTO THEIR ORDERED PLACES IN B2
421000
422000    ARRAY B1,B2,B1P,B2P[*] ;
423000    BEGIN
424000       REAL I1,I2,I3 ;
425000       LABEL L1,L2 ;
426000       FOR I1:=1 STEP 1 WHILE B1[I1]>0 DO
427000       BEGIN
428000          FOR I2:=1 STEP 1 UNTIL IBS DO IF B1[I1]=B2[I2] THEN GO TO L1 ;
429000          FOR I2:=1 STEP 1 UNTIL IBS DO
430000          IF B1[I1]>B2[I2] THEN
431000          BEGIN
432000             IBS := *+1 ;
433000             FOR I3:=IBS STEP -1 UNTIL I2+1 DO
434000             BEGIN B2[I3]:=B2[I3-1] ; B2P[I3]:=B2P[I3-1] ; END ;
435000             GO TO L2 ;
436000          END ;
437000          I3 := IBS := IBS+1 ;
438000          L2 : B2[I3] := B1[I1] ;
439000          B2P[I3] := B1P[I1] ;
440000          L1 : END ;
441000    END ;
442000
443000
444000
445000
446000
447000
448000    WRITE (LP,LYNE) ;
449000
450000    %        POSSIBLE NUCL MW CANDIDATES BY LOSSES FROM MW
451000
452000    WRITE (LP, < "POSSIBLE NUCL MW CANDIDATES ACCORDING TO LOSSES FROM M :",
453000    //,
454000    "  MW I  PR I     M         M-15      M-17      M-18      M-30      M-31
455000    M-32      M-35      M-36      M-61", /,
456000    "CAND I     I",X13,"(CH3)      (OH)     (H2O)    (H2C=O)   (CH2OH)   (CH
457000    3OH)   (OH+H2O) (H2O+H2O)  (30+31)",
458000    2( /,"      I      I")> ) ;
459000    IMW := 0 ;
460000    FOR I:=1 STEP 1 WHILE MASS[I]>=MASS[1]-61 DO
461000    %        FIND POSS MW CAND
462000    FOR V:=0,15,17,18,30,31,32,35,36,61 DO
463000    IF MW:=MASS[I]+V >=MASS[3] THEN
464000    BEGIN
465000       PR := CNT := 0 ;
466000       FOR W:=0,15,17,18,30,31,32,35,36,61 DO
467000       IF MW-W>0 THEN IF INT[MW-W]>0 THEN
468000       BEGIN PR:=*+PRFACT(MW-W) ; CNT:=*+1 ; END ;
469000       %        CRITERION : >=2 SUCH LOSSES
470000       IF CNT>=2 THEN
471000       BEGIN
472000          MNP := 1000 ;
473000          FLG := 0 ;
474000          %        IF MW CAND ALREADY INCLUDED
475000          FOR J:=1 STEP 1 UNTIL IMW DO
476000          BEGIN
477000             IF MCND[J]=MW THEN BEGIN MCNDPR[J]:=MAX( MCNDPR[J],PR ) ; FLG:=1 ; END ;
478000             MNP := MIN( MCNDPR[J],MNP ) ;
479000          END ;
480000          IF FLG=0 THEN
481000          IF IMW<30 THEN BEGIN IMW:=*+1 ; PLACECND( MW,PR,MCND,MCNDPR ) ; END
482000          ELSE
483000          WRITE (LP, < ">30 NUCL MW",/>) ;
484000       END ;
485000    END ;
486000    %        OUTPUT LINE OF LOSSES
487000    FOR I:=1 STEP 1 UNTIL IMW DO
488000    BEGIN
489000       J := -1 ;
490000       FOR V:=0,0,15,17,18,30,31,32,35,36,61 DO INS[J:=*+1]:=MCND[I]-V ;
491000       INS[0] := MCNDPR[I] ;
492000       OPUT(INS) ;
493000    END ;
494000    WRITE (LP,LYNE) ;
495000
```

```
496000    %         CLASS 5 : D RIBOSE TYPE SUGAR
497000
498000    WRITE (LP, <"CLASS 5 : D RIBOSE TYPE SUGAR", //> ) ;
499000    %         BASE CANDIDATES
500000    RIBOSE(30,44,60) ;
501000    FOR I:=1 STEP 1 UNTIL IBS DO
502000    BEGIN BRIB[I]:=BCND[I] ; BRIBPR[I]:=BCNDPR[I] ; END ;
503000
504000    %         CLASS 6 : 2' OME TYPE SUGAR
505000
506000    WRITE (LP, <"CLASS 6 : 2' OME TYPE SUGAR",//>) ;
507000    %         BASE CANDIDATES
508000    RIBOSE(30,58,74) ;
509000    FOR I:=1 STEP 1 UNTIL IBS DO
510000    BEGIN B2OM[I]:=BCND[I] ; B2OMPR[I]:=BCNDPR[I] ; END ;
511000
512000    %         CLASS 7 : 2' DEOXY TYPE SUGAR
513000
514000    WRITE (LP, <"CLASS 7 : 2' DEOXY TYPE SUGAR", //>) ;
515000    %         BASE CANDIDATES
516000    RIBOSE(30,28.0) ;
517000    FOR I:=1 STEP 1 UNTIL IBS DO
518000    BEGIN B2DO[I]:=BCND[I] ; B2DOPR[I]:=BCNDPR[I] ; END ;
519000
520000    %         COLLECT ALL BASE CAND TOGETHER, IN BCND
521000    %         AND FIND POSS MW BY COMBINING THEM WITH THE 3 COMMON SUGAR MASS
522000
523000    BCOLLECT(BRIB,BRIBPR,BCND,BCNDPR) ;
524000    BCOLLECT(B2OM,B2OMPR,BCND,BCNDPR) ;
525000    %         FIDDLE THE BASE PR
526000    FOR I:=1 STEP 1 UNTIL IBS DO BCNDPR[I] := BCNDPR[I]/I ;
527000    WRITE (LP, < "BASE CANDIDATES COLLECTED AND RE RANKED",///,
528000    *I4, //, *I4, />,
529000    IBS, FOR I:=1 STEP 1 UNTIL IBS DO BCND[I],
530000    IBS, FOR I:=1 STEP 1 UNTIL IBS DO BCNDPR[I] ) ;
531000    WRITE (LP,LYNE) ;
532000    WRITE (LP, < "LOSSES FROM POSSIBLE MW = B+133      =MW OF D RIBOSE SUGAR,
533000    ET AL) :"- />) ;
534000    LOSMW(133,M133,M133PR) ;
535000    WRITE (LP, < "LOSSES FROM POSSIBLE MW = B+147      =MW OF 2' OR 3' OME RI
536000    BOSE SUGAR, ET AL) :", />) ;
537000    LOSMW(147,M147,M147PR) ;
538000    WRITE (LP, < "LOSSES FROM POSSIBLE MW = B+117      = MW OF 2' OR 3' DEOXY
539000    RIBOSE SUGAR, ET AL) : "- />) ;
540000    LOSMW(117,M117,M117PR) ;
541000
542000    %         OCCAISIONAL  SUGAR IONS
543000
544000    WRITE (LP, <"OCCAISIONAL SUGAR IONS",//>) ;
545000    IF INT[133]>0 THEN WRITE (LP,<"133 INT=",I4,"  => SUG=133",/>,
546000    INT[133] ) ;
547000    IF INT[146]>0 THEN WRITE (LP,<"146 INT=",I4,"  => SUG=147",/>,
548000    INT[146] ) ;
549000    IF INT[117]>0 THEN WRITE (LP,<"117 INT=",I4,"  => SUG=117",/>,
550000    INT[117] ) ;
551000    WRITE (LP,LYNE) ;
552000 END ;
553000
554000

555000 PROCEDURE EXEC ;
556000 BEGIN
557000    ARRAY ORD,MW1,MW2,PRMW1,PRMW2[0:21], ORMAS,ORINT,O2INT,O2MAS[0:300] ;
558000    ARRAY BWT5,BPR5,CGLYC,CGLYCPR[0:60] ;
559000    REAL I,ORNOPE,REM1,REM2,COUNT,BIGST ,O2NOPE,J,MS ;
560000    LABEL L1 ;
561000
562000    WRITE (LP[SKIP 1] ) ;
563000    WRITE (LP, <13A6,/////>, CMMNT) ;
564000    REM1 := 0 ;
565000    FOR I:=0,1,2,3,4,5,6,7,8,9,10,11,12 DO
566000    IF NAME[I] NEQ "      " THEN REM1 := 1 ;
567000    IF REM1=1 THEN WRITE (LP, <13A6,/>, NAME ) ;
568000    REM1 := 0 ;
569000    FOR I:=0,1,2,3,4,5,6,7,8,9,10,11,12 DO
570000    IF CMMNTSOT[I] NEQ "      " THEN REM1:=1 ;
571000    IF REM1=1 THEN WRITE (LP, <13A6,/>, CMMNTSOT ) ;
572000    WRITE (LP, <"SPECTRUM NO.",I4,/>, ISPEC ) ;
573000
```

```
574000    %        KEEP COPY, AND REVERSE DATA IF NEC.
575000    IF REVDATA=1 THEN
576000    BEGIN
577000       FOR I:=1 STEP 1 UNTIL NOPE DO
578000       BEGIN ORMAS[I]:=MASS[NOPE-I+1] ; ORINT[I]:=INT[NOPE-I+1] ; END ;
579000       FOR I:=1 STEP 1 UNTIL NOPE DO
580000       BEGIN MASS[I]:=ORMAS[I] ; INT[I]:=ORINT[I] ; END ;
581000    END
582000    ELSE
583000    FOR I:=1 STEP 1 UNTIL NOPE DO
584000    BEGIN ORMAS[I]:=MASS[I] ; ORINT[I]:=INT[I] ; END ;
585000    ORNOPE := NOPE ;
586000    DO VECTORMODE (O2MAS,ORMAS,FOR NOPE+1)
587000    BEGIN O2MAS:=ORMAS ; INCREMENT O2MAS,ORMAS ; END ;
588000    DO VECTORMODE (O2INT,ORINT,FOR NOPE+1)
589000    BEGIN O2INT:=ORINT ; INCREMENT O2INT,ORINT ; END ;
590000
591000    %        REMOVE SMALL PEAKS (< CONST THROUGHOUT RANGE)
592000    FOR I:=1 STEP 1 UNTIL NOPE DO
593000    IF O2INT[I]<=MININT THEN
594000    BEGIN
595000       FOR J:=I STEP 1 UNTIL NOPE-1 DO
596000       BEGIN O2MAS[J]:=O2MAS[J+1] ; O2INT[J]:=O2INT[J+1] ; END ;
597000       I := *-1 ;
598000       NOPE := *-1 ;
599000    END ;
600000
601000    %        REMOVE MASS DEFICIENT PEAKS
602000    IF MASSDEF=1 THEN
603000    FOR I:=1 STEP 1 UNTIL NOPE DO IF O2MAS[I] MOD 1 >=0.5 THEN
604000    BEGIN
605000       FOR J:=I STEP 1 UNTIL NOPE-1 DO
606000       BEGIN O2MAS[J]:=O2MAS[J+1] ; O2INT[J]:=O2INT[J+1] ; END ;
607000       I:=*-1 ;
608000       NOPE:=*-1 ;
609000    END ;
610000
611000    %        INTEGERISE, REMOVE MULTIPLES, AND CONVERT TO INT[MASS[I]] FORM
612000    I := 1 ;
613000    FOR J:=1 STEP 1 UNTIL NOPE DO
614000    BEGIN
615000       BIGST := 0 ;
616000       IF INTEGER(O2MAS[J])=INTEGER(O2MAS[J+1]) THEN
617000       BEGIN
618000          L1 : BIGST := MAX(BIGST,O2INT[J],O2INT[J+1]) ;
619000          J := *+1 ;
620000          IF INTEGER(O2MAS[J])=INTEGER(O2MAS[J+1]) AND J<=NOPE THEN GO TO L1 ;
621000          MASS[I] := INTEGER(O2MAS[J]) ;
622000          INT[MASS[I]] := BIGST ;
623000       END
624000       ELSE
625000       BEGIN MASS[I]:=INTEGER(O2MAS[J]) ; INT[MASS[I]]:=O2INT[J] ; END ;
626000       I:=*+1 ;
627000    END ;
628000    NOPE:=I-1 ;
629000
630000    %        KEEP A COPY FOR MOLION
631000    BEGIN
632000       FOR I:=1 STEP 1 UNTIL NOPE DO
633000       BEGIN O2MAS[I]:=MASS[I] ; O2INT[I]:=INT[MASS[I]] ; END ;
634000       O2NOPE := NOPE ;
635000    END ;
636000
637000    %        REMOVE ISOTOPE PEAKS
638000    FOR I:=NOPE STEP -1 UNTIL 2 DO
639000    BEGIN
640000       INT[MASS[I]+1] := *-4.4@-4*MASS[I]*INT[MASS[I]] ;
641000       INT[MASS[I]+2] := *-((MASS[I]-25)*1.0@-7+4.0@-5)*MASS[I]*INT[MASS[I]] ;
642000       IF INT[MASS[I]+1]<0 THEN INT[MASS[I]+1]:=0 ;
643000       IF INT[MASS[I]+2]<0 THEN INT[MASS[I]+2]:=0 ;
644000    END ;
645000
646000    %        NORMALISE IF NEC.
647000    FOR I:=1 STEP 1 UNTIL NOPE DO BIGST:=MAX(INT[MASS[I]],BIGST) ;
648000    IF BIGST^=100 AND BIGST>0 THEN
649000    BEGIN
650000       BIGST := 100/BIGST ;
651000       FOR I:=1 STEP 1 UNTIL NOPE DO INT[MASS[I]]:=INT[MASS[I]]*BIGST ;
652000    END ;
653000
```

```
654000     %        REORDER MASS[I] + SET SMALL PEAKS TO 0
655000     FOR I:=1 STEP 1 UNTIL NOPE DO
656000     IF INT[MASS[I]]<=0.5*10**(2-MASS[I]/50) + 600/(MASS[I]+10) -2
657000     OR INT[MASS[I]]=0 THEN
658000     BEGIN
659000        REM1 := I ;
660000        FOR I:=*+1 STEP 1 UNTIL NOPE-1 DO
661000        IF INT[MASS[I]]<=0.5*10**(2-MASS[I]/50) + 600/(MASS[I]+10) -2
662000        OR INT[MASS[I]]=0 THEN
663000        BEGIN
664000           REM2 := I ;
665000           COUNT := *+1 ;
666000           INT[MASS[REM1]] := 0 ;
667000           DO BEGIN MASS[REM1]:=MASS[REM1+COUNT] ; REM1:=*+1 ; END UNTIL
668000           REM1+COUNT > REM2 ;
669000           REM1 := *-1 ;
670000        END ;
671000        REM2 := I ;
672000        COUNT := *+1 ;
673000        INT[MASS[REM1]] := 0 ;
674000        DO BEGIN MASS[REM1]:=MASS[REM1+COUNT] ; REM1:=*+1 ; END UNTIL
675000        REM1+COUNT > REM2 ;
676000        NOPE := *-COUNT ;
677000        IF INT[MASS[NOPE]]<=0.5*10**(2-MASS[NOPE]/50) + 600/(MASS[NOPE]+10) -2
678000        OR INT[MASS[NOPE]]=0 THEN BEGIN INT[MASS[NOPE]]:=0 ; NOPE:=*-1 ; END ;
679000     END ;
680000
681000     %        INPUT CHECK
682000     FOR I:=1 STEP 1 UNTIL NOPE-1 DO
683000     IF MASS[I]-MASS[I+1]>1 THEN
684000     FOR MS:=MASS[I]-1 STEP -1 WHILE MS-MASS[I+1]>0 DO INT[MS]:=0 ;
685000     FOR MS:=MASS[NOPE]-1 STEP -1 UNTIL 1 DO INT[MS]:=0 ;
686000     FOR I:=2 STEP 1 UNTIL ORNOPE DO
687000     IF INTEGER ORMAS[I])>INTEGER ORMAS[I-1]) THEN
688000     WRITE (LP, <"ORIGINAL INPUT ERROR MASS",I3,///>, I) ;
689000     FOR I:=2 STEP 1 UNTIL NOPE DO
690000     IF MASS[I]>=MASS[I-1] THEN
691000     WRITE (LP, <"INPUT ERROR MASS",I3,///>,I) ;
692000
693000     %        OUTPUT
694000     IF SPLIST=1 THEN
695000     BEGIN
696000        IF MASSDEF=1 THEN
697000        WRITE (LP, <X11,"ORIGINAL SPECT.",X8,"NORMALISED,WITH LOW ,ISOTOPE,AND M
698000        ASS DEFICIENT PEAKS REMOVED.",////,
699000        X12,"MASS",X6,"INT.",X17,"MASS",X6,"INT.",//> )
700000        ELSE
701000        WRITE (LP, <X11,"ORIGINAL SPECT.",X8,"NORMALISED,WITH LOW AND ISOTOPE PE
702000        AKS REMOVED.",////,X12,"MASS",X6."INT.",X17,"MASS",X6,"INT "-//> ) ;
703000        FOR I:=1 STEP 1 UNTIL ORNOPE DO
704000        IF I<=NOPE THEN
705000        WRITE (LP, <I3,F13.5,F10.2,X8,I3,2F10.2,/>, I,ORMAS[I],ORINT[I],I,
706000        MASS[I],INT[MASS[I]] )
707000        ELSE
708000        WRITE (LP, <I3,F13.5,F10.2,/>, I,ORMAS[I],ORINT[I] ) ;
709000     END
710000     ELSE
711000     IF SPLIST=2 THEN
712000     BEGIN
713000        WRITE (LP,<"ORIGINAL SPECT.",//>) ;
714000        WRITE (LP, < 10 (" ",I3,",",I3,")  " ),/ >,
715000        FOR I:=1 STEP 1 UNTIL ORNOPE DO [ ORMAS[I],ORINT[I] ] ) ;
716000        IF MASSDEF=1 THEN
717000        WRITE (LP, <///,"NORMALISED,WITH LOW,ISOTOPE AND MASS DEFICIENT PEAKS RE
718000        MOVED",//>)
719000        ELSE
720000        WRITE (LP, <///,"NORMALISED WITH LOW AND ISOTOPE PEAKS REMOVED",//>) ;
721000        WRITE (LP, < 10 (" ",I3,",",I3,")  " ),/ >,
722000        FOR I:=1 STEP 1 UNTIL NOPE DO [ MASS[I],INT[MASS[I]] ] ) ;
723000        WRITE (LP, <//>) ;
724000     END ;
725000     IF ISHRT=0 THEN
726000     WRITE (LP, <"INPUT DATA :",//,
727000     "REVDATA=",I2," SUGANAL=",I2,  " MWSEARCH=",I2, " MASSDEF=",I2,
728000     " PAMI=",I2," SPLIST=",I2,// ,
729000     "ISHRT=",I2," MININT=",I3,  " CL124=",I2," CL567=",I2,/>,
730000     REVDATA,SUGANAL,MWSEARCH,  MASSDEF,PAMI,SPLIST,ISHRT,MININT,CL124,
731000     CL567 ) ;
732000     WRITE (LP, <//////> ) ;
733000
```

```
734000    %        MW ROUTINE
735000    IF MWSEARCH=1 THEN
736000    BEGIN
737000       IF PAMI=0 THEN
738000       BEGIN
739000          FOR I:=1 STEP 1 UNTIL O2NOPE DO
740000          BEGIN ORMAS[I-1]:=O2MAS[I] ; ORINT[I-1]:=O2INT[I] ; END ;
741000          ORNOPE := O2NOPE ;
742000       END
743000       ELSE
744000       IF PAMI=1 THEN
745000       BEGIN
746000          FOR I:=1 STEP 1 UNTIL NOPE DO
747000          BEGIN ORMAS[I-1]:=MASS[I] ; ORINT[I-1]:=INT[MASS[I]] ; END ;
748000          ORNOPE := NOPE ;
749000       END ;
750000       FOR I:=0 STEP 1 UNTIL ORNOPE-1 DO BIGST:=MAX(ORINT[I],BIGST) ;
751000       IF BIGST>0 AND BIGST^=500 THEN
752000       BEGIN
753000          BIGST := 500/BIGST ;
754000          FOR I:=0 STEP 1 UNTIL ORNOPE-1 DO ORINT[I]:=INTEGER ORINT[I]*BIGST) ;
755000       END ;
756000       IF ORNOPE>4 THEN
757000       BEGIN
758000          MOLION(ORMAS,ORINT,ORNOPE,ISHRT) ;
759000          WRITE (LP,<//>) ;
760000       END ;
761000    END ;
762000
763000    %        CARBOHYDRATE ANALYSIS
764000    IF SUGANAL=1 THEN
765000    BEGIN
766000       IF CL124=1 THEN CLASSGT ;
767000       IF CL567=1 THEN CLASSGTB(BWT5,BPR5) ;
768000       IF CL8=1 THEN CLASSGTC(CGLYC,CGLYCPR) ;
769000
770000       FOR I:=1 STEP 1 UNTIL NOCL DO
771000       IF PR[I]>0 THEN
772000       CASE I OF
773000       BEGIN
774000          1 : CLSPEC1(MW1,PRMW1) ;
775000          2 : CLSPEC2(MW2,PRMW2) ;
776000          4 : CLSPEC4 ;
777000       END ;
778000       IF BWT5[1]>0 THEN CLSPEC5(BWT5,BPR5) ;
779000
780000       %        ORDER CLASS PRS
781000       %        ORD[1]
782000       BIGST := PR[1] ;
783000       ORD[1] := 1 ;
784000       FOR I:=2 STEP 1 UNTIL NOCL DO
785000       IF BIGST<PR[I] THEN BEGIN BIGST:=PR[I] ; ORD[1]:=I ; END ;
786000       %        ORD[2]
787000       IF PR[ORD[1]]=0 THEN ORD[1]:=0 ELSE
788000       FOR I:=1 STEP 1 UNTIL NOCL DO
789000       IF I^=ORD[1] THEN
790000       BEGIN
791000          BIGST:=PR[I] ; ORD[2]:=I ;
792000          FOR I:=*+1 STEP 1 UNTIL NOCL DO
793000          IF I^=ORD[1] THEN
794000          IF BIGST<PR[I] THEN BEGIN BIGST:=PR[I] ; ORD[2]:=I ; END ;
795000          %        ORD[3]
796000          IF PR[ORD[2]]=0 THEN ORD[2]:=0 ELSE
797000          BEGIN
798000             FOR I:=1 STEP 1 UNTIL NOCL DO
799000             IF I^=ORD[1] AND I^=ORD[2] THEN
800000             BEGIN
801000                BIGST:=PR[I] ; ORD[3]:=I ;
802000                FOR I:=*+1 STEP 1 UNTIL NOCL DO
803000                IF I^=ORD[1] AND I^=ORD[2] THEN
804000                IF BIGST<PR[I] THEN BEGIN BIGST:=PR[I] ; ORD[3]:=I ; END ;
805000             END ;
```

```
806000              %         ORD[4]
807000              IF PR[ORD[3]]=0 THEN ORD[3]:=0 ELSE
808000              BEGIN
809000                 FOR I:=1 STEP 1 UNTIL NOCL DO
810000                 IF I^=ORD[1] AND I^=ORD[2] AND I^=ORD[3] THEN
811000                 BEGIN
812000                    BIGST:=PR[I] ; ORD[4]:=I ;
813000                    FOR I:=*+1 STEP 1 UNTIL NOCL DO
814000                    IF I^=ORD[1] AND I^=ORD[2] AND I^=ORD[3] THEN
815000                    IF BIGST<PR[I] THEN BEGIN BIGST:=PR[I] ; ORD[4]:=I ; END ;
816000                 END ;
817000                 %         ORD[5]
818000                 IF PR[ORD[4]]=0 THEN ORD[4]:=0 ELSE
819000                 BEGIN
820000                    FOR I:=1 STEP 1 UNTIL NOCL DO
821000                    IF I^=ORD[1] AND I^=ORD[2] AND I^=ORD[3] AND I^=ORD[4] THEN
822000                    BEGIN
823000                       BIGST:=PR[I] ; ORD[5]:=I ;
824000                       FOR I:=*+1 STEP 1 UNTIL NOCL DO
825000                       IF I^=ORD[1] AND I^=ORD[2] AND I^=ORD[3] AND I^=ORD[4] THEN
826000                       IF BIGST<PR[I] THEN BEGIN BIGST:=PR[I] ; ORD[5]:=I ; END ;
827000                    END ;
828000                    IF PR[ORD[5]]=0 THEN ORD[5]:=0 ;
829000                 END ;
830000              END ;
831000           END ;
832000        END ;
833000        WRITE (LP, <"RANKED CLASS PROBABILITIES (MAX 5)",/>) ;
834000        WRITE (LP, <"CLASS",5I5>, FOR I:=1 STEP 1 WHILE I<=5 AND ORD[I]>0 DO
835000        ORD[I] ) ;
836000        WRITE (LP) ;
837000        WRITE (LP, <"PR    ",5I5>, FOR I:=1 STEP 1 WHILE I<=5 AND ORD[I]>0 DO
838000        PR[ORD[I]] ) ;
839000     END ;
840000 END ;
841000

842000 PROCEDURE MWLOSS(M,P,N,MAXLOSS,PM,SM) ;
843000 ARRAY M,P,SM,PM[*] ;
844000 INTEGER MAXLOSS,N ;
845000 BEGIN
846000     REAL G,H,I,J,K,L,NPL,NSL,FLG ;
847000     ARRAY PL      ,SL,LVL[0:20] ;
848000     FOR K:=1 STEP 1 UNTIL N DO
849000
850000     %         PRIMARY LOSSES
851000     BEGIN
852000        J := 0 ;
853000        FOR G:=1 STEP 1 WHILE L:=PM[G]>0 DO
854000        IF M[K]-L>0 THEN IF INT[M[K]-L]>0 THEN
855000        BEGIN J:=*+1 ; PL[J]:=M[K]-L ; P[K]:=*+0.001*PL[J]*INT[PL[J]] ; END ;
856000        NPL := J ;
857000
858000        %         FIRST SECONDARY LOSSES
859000        J := 0 ;
860000        FOR  I:=1 STEP 1 UNTIL NPL DO
861000        FOR G:=1 STEP 1 WHILE L:=SM[G]>0 DO
862000        IF PL[I]-L>0 THEN IF INT[PL[I]-L]>0 THEN
863000        BEGIN
864000           FOR H:=1 STEP 1 UNTIL J DO IF PL[I]-L=SL[H] THEN FLG:=1 ;
865000           IF FLG=0 THEN
866000           BEGIN J:=*+1 ; SL[J]:=PL[I]-L ; P[K]:=*+0.001*SL[J]*INT[SL[J]] ; END ;
867000           FLG:=0 ;
868000        END ;
869000        NSL := J ;
870000        FOR I:=1 STEP 1 UNTIL NSL DO LVL[I] := 2 ;
871000
```

```
872000        %         SUBSEQUENT SECONDARY LOSSES
873000        FOR I:=1 STEP 1 UNTIL NSL DO
874000        IF LVL[NSL]<MAXLOSS THEN
875000        FOR J:=1 STEP 1 WHILE L:=SM[J]>0 DO
876000        IF SL[I]-L>0 THEN IF INT[SL[I]-L]>0 THEN
877000        BEGIN
878000           FOR H:=1 STEP 1 UNTIL NSL DO IF SL[I]-L=SL[H] THEN G:=1 ;
879000           IF G=0 THEN
880000           BEGIN
881000              NSL := *+1 ;
882000              SL[NSL] := SL[I]-L ;
883000              LVL[NSL] := LVL[I]+1 ;
884000              P[K] := *+0.001*SL[NSL]*INT[SL[NSL]] ;
885000           END ;
886000           G := 0 ;
887000        END ;
888000
889000        IF NPL>0 OR NSL>0 THEN
890000        BEGIN
891000           G:=0 ;
892000           DO G:=*+1 UNTIL PM[G]=0 ;
893000           H:=0 ;
894000           DO H:=*+1 UNTIL SM[H]=0 ;
895000           WRITE (LP, <"LOSSES OF",*I4,"  THEN",*I4, "  FROM MW",I2, " =",I5,
896000           "  GIVE",//, *I5,*I5,//,"  HENCE PR INCR TO",I5,/> ,
897000           G-1,FOR I:=1 STEP 1 UNTIL G-1 DO PM[I],
898000           H-1,FOR I:=1 STEP 1 UNTIL H-1 DO SM[I],            K,M[K],
899000           NPL,FOR I:=1 STEP 1 UNTIL NPL DO PL[I],
900000           NSL,FOR I:=1 STEP 1 UNTIL NSL DO SL[I],      P[K] )
901000        END
902000        ELSE
903000        WRITE (LP, <"NO REGULAR PRIMARY LOSSES FROM MW",I2," =",I5,/>, K,M[K]) ;
904000        FOR H:=0,1,2,3,4 DO PL[H] := 0 ;
905000        FOR H:=0 STEP 1 UNTIL 20 DO SL[H]:=0 ;
906000     END ;
907000     WRITE (LP, <//>) ;
908000  END ;
909000

910000  PROCEDURE(ORDMW M,P,N) ;
911000  ARRAY M,P[*] ;
912000  INTEGER N ;
913000  BEGIN
914000     REAL I,SWOP ;
915000     FOR I :=1 STEP 1 UNTIL N-1 DO
916000     IF P[I]<P[I+1] THEN
917000     BEGIN
918000        SWOP:=P[I+1] ; P[I+1]:=P[I] ; P[I]:=SWOP ;
919000        SWOP:=M[I+1] ; M[I+1]:=M[I] ; M[I]:=SWOP ;
920000        I := MAX(0,I-3) ;
921000     END ;
922000     WRITE (LP, <X3,*I6,/>, N,FOR I:=1 STEP 1 UNTIL N DO I ) ;
923000     WRITE(LP,<" MW",*I6,/>,N,FOR I:=1 STEP 1 UNTIL N DO M[I] ) ;
924000     WRITE(LP,<" PR",*I6,/>,N,FOR I:=1 STEP 1 UNTIL N DO P[I] ) ;
925000  END ;
926000

927000  PROCEDURE HPCLSTR(X) ; *

928000  PROCEDURE X ;
929000  BEGIN
930000     HP := MASS[1] ;
931000     X ;
932000     IF MASS[1]-MASS[2]<4 THEN
933000     BEGIN
934000        HP := MASS[2] ;
935000        X ;
936000        IF MASS[2]-MASS[3]<4 THEN
937000        BEGIN
938000           HP := MASS[3] ;
939000           X ;
940000        END ;
941000     END ;
942000  END ;
```

2-METHYLADENOSINE SUG 133 BASE 148 MW281 REF ANB08

SPECTRUM NO. 1

ORIGINAL SPECT.

281, 1	264, 1	251, 2	250, 1	246, 1	239, 1	234, 1	233, 1	232, 1	231, 1
216, 1	213, 1	211, 1	208, 1	206, 1	203, 1	202, 1	200, 1	199, 1	192, 6
178, 15	166, 1	150, 15	149, 20	133, 2	109, 3	108, 4	81, 2	74, 30	59, 40

NORMALISED,WITH LOW AND ISOTOPE PEAKS REMOVED

281, 3	264, 3	251, 5	250, 3	246, 3	239, 3	234, 2	233, 2	232, 2	231, 3
216, 3	213, 2	211, 3	208, 2	206, 3	203, 2	202, 2	200, 2	199, 3	192, 15
178, 38	166, 3	150, 34	149, 50	133, 5	109, 7	108, 10	74, 75	59,100	

INPUT DATA :

REVDATA= 0 SUGANAL= 1 MWSEARCH= 1 MASSDEF= 0 PAMI= 1 SPLIST= 2

ISHRT= 0 MININT= 0 CL124= 0 CL567= 1

```
 59   500   74   375  108    50  109    35  133    25  149   250
150   171  166    13  178   188  192    75  199    13  200    11
202    12  206    13  208    12  211    13  213    12  216    13
231    13  232    11  233    11  234    11  239    13  246    13
250    13  251    24  264    13  281    13

NEVPR=    0  NODPR=    1  SEVPR=      1.  SODPR=      0.

HYESTM =  281  BASEPR =    0  NITPKS =    0  ARING =    0

IVSHRT=    1  ISHRT=    0  MSTART=  280  NTOP=   3  MOLFLG=    0

NR=   4  NR6=    4  NOUT=   27  NMOL=    0  ISOMAS=  281

  16   594    18   110    26   137    28   586    30   249
  32    72    34   499    36    25    42   517    44    51
  46    24    48    51    54   287    56   209    58   360
  62   288    68   201    70   238    72   289    74   900
  76   546    78    38    80    63    82   445    84   412
  86   201    90  1061    92   449    94    62    96   184
  98   127   100   282   102   322   104   610   106    38
 108   113   114   184
  15  1521    17   643    19   172    25   137    27   610
  29   712    31   168    33   887    35   698    41  1109
  43   412    45   300    47   190    49   770    53   287
  55   408    57   671    59  1361    61   673    63   183
  67   524    69   498    71   289    73  1176    75  1208
  77    38    79    63    81   629    83  1003    85   238
  89  1245    91  1229    93   109    95   184    97   438
  99   365   101   780   103   673   105   100   107   674
 109    35   113   222   115   263

CANDIDATE MOLIONS AND THEIR RANKINGS

***      281   **      100    *** BEST CANDIDATE

***      282   **       77    ***

***      296   **       64    ***

***      297   **        2    ***

***      299   **       69    ***

***      300   **       66    ***

***      308   **       12    ***

***      309   **        9    ***
```

***	310	**	44	***
***	311	**	29	***
***	312	**	53	***
***	313	**	60	***
***	322	**	14	***
***	323	**	37	***
***	324	**	48	***
***	325	**	29	***
***	326	**	32	***
***	328	**	30	***
***	336	**	40	***
***	338	**	31	***
***	339	**	40	***
***	341	**	22	***
***	342	**	32	***
***	352	**	24	***
***	353	**	25	***
***	354	**	58	***
***	366	**	8	***
***	380	**	3	***
***	381	**	2	***
***	382	**	14	***
***	383	**	3	***
***	394	**	2	***
***	395	**	1	***
***	396	**	2	***

POSSIBLE NUCL MW CANDIDATES ACCORDING TO LOSSES FROM M :

MW CAND	PR	M	M-15 CH3	M-17 OH	M-18 H2O	M-30 H2C=O	M-31 CH2OH	M-32 CH3OH	M-35 OH+H2O	M-36 H2O+H2O	M-61 30+31
312	2						281, 3				251, 5
311	1					281, 3					250, 3
300	1									264, 3	239, 3
299	1				281, 3				264, 3		
296	1		281, 3					264, 3			
295	1						264, 3				234, 2
294	1					264, 3					233, 2
286	2								251, 5	250, 3	
282	3				264, 3		251, 5	250, 3		246, 3	
281	4	281, 3		264, 3		251, 5	250, 3		246, 3		
277	1						246, 3				216, 3
274	1								239, 3		213, 2
270	1						239, 3			234, 2	
269	3				251, 5	239, 3			234, 2	233, 2	208, 2
268	3			251, 5	250, 3				233, 2	232, 2	
267	2			250, 3					232, 2	231, 3	206, 3
266	2		251, 5					234, 2	231, 3		
265	2		250, 3				234, 2	233, 2			
264	3	264, 3			246, 3	234, 2	233, 2	232, 2			203, 2
263	3			246, 3		233, 2	232, 2	231, 3			202, 2
262	1					232, 2	231, 3				
261	2		246, 3			231, 3					200, 2
252	1				234, 2					216, 3	
251	3	251, 5		234, 2	233, 2				216, 3		

CLASS 5 : D RIBOSE TYPE SUGAR

CHARACTERISTIC IONS FOR B :

BASE CAND	PR	B	B+1	B+2	B+14	B+15	B+28 B+44-16	B+30	B+44	B+58 B+44+14	B+60	B+74 B+60+14
148	23		149, 50	150, 34				178, 38	192, 15	206, 3	208, 2	
106	7			108, 10					150, 34		166, 3	

****

CLASS 6 : 2' OME TYPE SUGAR

CHARACTERISTIC IONS FOR B :

BASE CAND	PR	B	B+1	B+2	B+14	B+15	B+28 B+44-16	B+30	B+44	B+58 B+44+14	B+60	B+74 B+60+14
176	8			178, 38				206, 3		234, 2		250, 3
148	23		149, 50	150, 34				178, 38	192, 15	206, 3	208, 2	

****

CLASS 7 : 2' DEOXY TYPE SUGAR

CHARACTERISTIC IONS FOR B :

BASE CAND	PR	B	B+1	B+2	B+14	B+15	B+28 B+44-16	B+30	B+44	B+58 B+44+14	B+60	B+74 B+60+14
148	23		149, 50	150, 34				178, 38	192, 15	206, 3	208, 2	

****

BASE CANDIDATES COLLECTED AND RE RANKED

176.148 106

LOSSES FROM POSSIBLE MW = B+133 =MW OF D RIBOSE SUGAR, ET AL :

MW CAND	FROM B	PR	M	M-15 CH3	M-17 OH	M-18 H2O	M-30 H2C=O	M-31 CH2OH	M-32 CH3OH	M-35 OH+H2O	M-36 H2O+H2O	M-61 30+31
281	148	4	281, 3		264, 3		251, 5	250, 3		246, 3		

LOSSES FROM POSSIBLE MW = B+147 =MW OF 2' OR 3' OME RIBOSE SUGAR, ET AL :

MW CAND	FROM B	PR	M	M-15 CH3	M-17 OH	M-18 H2O	M-30 H2C=O	M-31 CH2OH	M-32 CH3OH	M-35 OH+H2O	M-36 H2O+H2O	M-61 30+31
295	148	1						264, 3				234, 2
253	106	3										192, 15

LOSSES FROM POSSIBLE MW = B+117 = MW OF 2' OR 3' DEOXY RIBOSE SUGAR, ET AL :

MW CAND	FROM B	PR	M	M-15 CH3	M-17 OH	M-18 H2O	M-30 H2C=O	M-31 CH2OH	M-32 CH3OH	M-35 OH+H2O	M-36 H2O+H2O	M-61 30+31
293	176	1										232, 2
265	148	2		250, 3				234, 2	233, 2			

OCCAISIONAL SUGAR IONS

133 INT= 5 => SUG=133

CHARACTERISTIC C GLYCOSIDE IONS :

BASE CAND	PR	B+14	B+15	B+28	B+30 ====	B+44	B+58	B+60	B+74
120	12				150, 34		178, 38		
119	8	133, 5			149, 50				

****

URACIL TYPE BASE ASSUMPTION :

CHARACTERISTIC IONS 69, FROM B+1 = 112 :

BASE I ORIG I NEW I	B+1-43
CAND I PR I PR I	HNCO
I I I	
I I I	

****

CYTOSINE TYPE BASE ASSUMPTION :

CHARACTERISTIC IONS 95, 83, 69,151, FROM B+1 = 111 :

BASE I ORIG I NEW I	B+1-16	B+1-28	B+1-42	B+41
CAND I PR I PR I	NH2	CO	NCO	B-CO-CH
I I I				
I I I				
176 I 8 I 16 I		149, 50		
I I I				
148 I 12 I 12 I	133, 5			
I I I				

****

GUANINE TYPE BASE ASSUMPTION :

CHARACTERISTIC IONS 135,134,110,109,107, 81, 79, FROM B+1 = 151 : 109, 7

BASE I ORIG I NEW I	B+1-16	B+1-17	B+1-41	B+1-42	B+1-44	B+1-70	B+1-72
CAND I PR I PR I	NH2	NH3	HNCN	NCO	NH2+CO	NH2+HCN +HCN	NH2+CO +CO
I I I							
I I I							
176 I 8 I 9 I					133, 5		
I I I							
148 I 12 I 13 I	133, 5		108, 10				
I I I							

****

ADENINE TYPE BASE ASSUMPTION :

CHARACTERISTIC IONS 120,119,108, 81, FROM B+1 = 135 : 108, 10

BASE I ORIG I NEW I	B+1-15	B+1-16	B+1-27	B+1-54
CAND I PR I PR I	NH	NH2	HCN	HCN+HCN
I I I				
I I I				
176 I 8 I 13 I			150, 34	
I I I				
148 I 12 I 12 I		133, 5		
I I I				

****

Program KNNCLASSIF (Host)

```
1000 %     K NEAREST NEIGHBOUR AND DISTANCE FROM MEAN CLASSIFS
2000
3000 BEGIN
4000    FILE TAPEDATA(TITLE= "TAPEA123/SPECT/NUCL125.", KIND=DISK) ;
5000    FILE LP(KIND=PRINTER) , CR(KIND=READER) ,
6000    CP(KIND=PUNCH,TITLE="CHEM175.") ;
7000    ARRAY DATA[0:126,0:83], CLASS[0:126], MZ82[0:83] ;
8000    REAL CATLBL ;

9000    PROCEDURE CARDINPUT(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q) ;
10000   REAL A,B,C,D,E,F,G,H,I,J,K,L,M ; ARRAY N,O,P,Q[*] ; EXTERNAL ;

11000   PROCEDURE TAPEINPUT(A,B,C,D,E,F,G,H,I,J,K,L) ;
12000   REAL A,B,C,D,E,F,G,H,I,J,K ; ARRAY L[*] ; EXTERNAL ;

13000   PROCEDURE ARRANG(A,B,C,D,E,F,G) ; ARRAY A[*] ; REAL B,C,D,E,F,G ;
14000   EXTERNAL ;

15000   PROCEDURE AUTOSCALE(A,B,C,D) ; REAL A,B ; ARRAY C,D[*] ; EXTERNAL ;

16000   PROCEDURE REGEN(A,B,C,D) ; REAL A,B ; ARRAY C,D[*] ; EXTERNAL ;

17000   PROCEDURE EXEC(A,B,C,D,E,F,G,H,I,J) ; REAL A,B,C,D,E,F,G,H ;
18000   ARRAY I,J[*] ; EXTERNAL ;

19000   PROCEDURE DISTMEAN(A,B,C,D,E,F,G,H,I) ;
20000   REAL A,B,C,D,E,F,G,I ; ARRAY H[*] ; EXTERNAL ;

21000   PROCEDURE KNN(A,B,C,D,E,F,G,H,I) ;
22000   REAL A,B,C,D,E,F ; ARRAY G,H,I[*] ; EXTERNAL ;
23000   REAL IBCL,IELM,ISUG,ICLS,NCLS,NSBT,NELM,NRNGE1,NRNGE2,SGRMS,
24000   NICLS,NUM24,NUM12,I,NMEMTR,NNONTR,NMEM20,NNON20,NMEM49,NNON49 ;
25000   ARRAY AMU24[0:30], AMU12[0:15], CLSMEM[0:100], IPRSET[0:50] ;
26000   REAL T1,T2 ;
27000
28000
29000
30000   %     82 MOST POP M/Z VALUES
31000   FILL MZ82 WITH 0,108,109,110,111,112,113,114,115,117,119,120,121,125,
32000   126,127,133,134,135,136,139,140,141,146,148,149,151,152,154,155,160,162,
33000   163,164,165,166,168,169,170,171,176,177,178,179,180,183,185,190,191,192,
34000   193,194,200,201,202,206,207,208,209,211,218,219,220,221,225,226,227,228,
35000   232,240,248,249,250,251,257,258,266,267,268,269,280,281,316 ;
36000
37000   %     INPUT CONTROL PARAM
38000   CARDINPUT(IBCL,IELM,ISUG,ICLS,NCLS,NSBT,NELM,NRNGE1,NRNGE2,SGRMS,
39000   NICLS,NUM24,NUM12,AMU24,AMU12,CLSMEM,IPRSET) ;
40000
41000   %     INPUT 125 NUCL SPECT
42000   T1:=TIME(2) ;
43000   TAPEINPUT(IBCL,IELM,ISUG,ICLS,NCLS,NSBT,NELM,NRNGE1,NRNGE2,SGRMS,NICLS,
44000   CLSMEM) ;
45000   T2:=(TIME(2)-T1)/60 ;"
46000   WRITE (LP,</,"TAPEINPUT TIME",I5,/>,T2) ;
47000
48000   %     ORDER AND COUNT TR AND PR SETS
49000   T1:=TIME(2) ;
50000   ARRANG(IPRSET,NMEMTR,NNONTR,NMEM20,NNON20,NMEM49,NNON49) ;
51000   T2:=(TIME(2)-T1)/60 ;
52000   WRITE (LP,</,"ARRANG    TIME",I5,/>,T2) ;
53000
54000   %     KNN AND DIST FROM MEAN CLASSIFS
55000   EXEC(NMEMTR,NNONTR,NMEM20,NNON20,NMEM49,NNON49,NUM24,NUM12,AMU24,AMU12);
56000 END .
```

Program KNNCLASSIF (Procedures)

```
 1000 [
 2000 FILE TAPEDATA(TITLE= "TAPEA123/SPECT/NUCL125.", KIND=DISK) ;
 3000 FILE LP(KIND=PRINTER) , CR(KIND=READER) ,
 4000 CP(KIND=PUNCH,TITLE="CHEM175.") ;
 5000 ARRAY DATA[0:126,0:83], CLASS[0:126], MZ82[0:83] ;
 6000 REAL CATLBL ;
 7000

 8000 PROCEDURE CARDINPUT(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q) ;
 9000 REAL A,B,C,D,E,F,G,H,I,J,K,L,M ; ARRAY N,O,P,Q[*] ; EXTERNAL ;

10000 PROCEDURE TAPEINPUT(A,B,C,D,E,F,G,H,I,J,K,L) ;
11000 REAL A,B,C,D,E,F,G,H,I,J,K ; ARRAY L[*] ; EXTERNAL ;

12000 PROCEDURE ARRANG(A,B,C,D,E,F,G) ; ARRAY A[*] ; REAL B,C,D,E,F,G ;
13000 EXTERNAL ;

14000 PROCEDURE AUTOSCALE(A,B,C,D) ; REAL A,B ; ARRAY C,D[*] ; EXTERNAL ;

15000 PROCEDURE REGEN(A,B,C,D) ; REAL A,B ; ARRAY C,D[*] ; EXTERNAL ;

16000 PROCEDURE EXEC(A,B,C,D,E,F,G,H,I,J) ; REAL A,B,C,D,E,F,G,H ;
17000 ARRAY I,J[*] ; EXTERNAL ;

18000 PROCEDURE DISTMEAN(A,B,C,D,E,F,G,H,I) ;
19000 REAL A,B,C,D,E,F,G,I ; ARRAY H[*] ; EXTERNAL ;

20000 PROCEDURE KNN(A,B,C,D,E,F,G,H,I) ;
21000 REAL A,B,C,D,E,F ; ARRAY G,H,I[*] ; EXTERNAL ;
22000 ]
23000
24000
25000
26000

27000 PROCEDURE CARDINPUT(IBCL,IELM,ISUG,ICLS,NCLS,NSBT,NELM,NRNGE1,NRNGE2,
28000
29000 %      INPUT CONTROL PARAM AND OUTPUT AS CHECK
30000
31000 SGRMS,NICLS,NUM24,NUM12,AMU24,AMU12,CLSMEM,IPRSET) ;
32000 REAL IBCL,IELM,ISUG,ICLS,NCLS,NSBT,NELM,NRNGE1,NRNGE2,
33000 SGRMS,NICLS,NUM24,NUM12 ;
34000 ARRAY AMU24,AMU12,CLSMEM,IPRSET[*] ;
35000 BEGIN
36000    ARRAY TITLE[0:13] ;
37000    REAL I ;
38000    READ (CR,<13A6,A2>,TITLE) ;
39000    READ(CR,<A6>,CATLBL) ;
40000    READ (CR,<80I1>,IBCL,IELM,ISUG,ICLS) ;
41000    READ (CR,<26I3>, NCLS,NSBT,NELM,NRNGE1,NRNGE2,SGRMS,NICLS,NUM24,NUM12) ;
42000    READ (CR,<20I4>, FOR I:=1 STEP 1 UNTIL NUM24 DO AMU24[I]) ;
43000    READ (CR,<20I4>, FOR I:=1 STEP 1 UNTIL NUM12 DO AMU12[I]) ;
44000    READ (CR,<20I4>, FOR I:=1 STEP 1 UNTIL NICLS DO CLSMEM[I]) ;
45000    READ (CR,<26I3>,FOR I:=1 STEP 1 UNTIL 49 DO IPRSET[I]) ;
46000    WRITE (LP,<13A6,A2>,TITLE) ;
47000    WRITE (LP,</,"SHORT TITLE   ",A6>,CATLBL) ;
48000    WRITE (LP,<///"INPUT PARAMETERS",//,"  IBCL  IELM  ISUG  ICLS",//,
49000    20I6>, IBCL,IELM,ISUG,ICLS) ;
50000    WRITE (LP,<//,"  NCLS  NSBT  NELM RNGE1 RNGE2 SGRMS NICLS NUM24 NUM12",
51000    //,20I6>,
52000    NCLS,NSBT,NELM,NRNGE1,NRNGE2,SGRMS,NICLS,NUM24,NUM12) ;
53000    WRITE (LP,<//, "AMU24" ,//,30I4>, FOR I:=1 STEP 1 UNTIL NUM24 DO
54000    AMU24[I]) ;
55000    WRITE (LP,<//, "AMU12" ,//,30I4>, FOR I:=1 STEP 1 UNTIL NUM12 DO
56000    AMU12[I]) ;
57000    WRITE (LP,<//,"CLSMEM",//,(30I4)>, FOR I:=1 STEP 1 UNTIL NICLS DO
58000    CLSMEM[I]) ;
59000    WRITE (LP,<//, "PRSET",//,(30I4)>,FOR I:=1 STEP 1 UNTIL 49 DO
60000    IPRSET[I]) ;
61000 END ;
62000
```

```
63000 PROCEDURE TAPEINPUT(IBCL,IELM,ISUG,ICLS,NCLS,NSBT,NELM,NRNGE1,
64000 NRNGE2,SGRMS,NICLS,CLSMEM);
65000
66000       INPUT 125 SPECT FROM DATA FILE AND ASSIGN CLASS
67000
68000 REAL IBCL,IELM,ISUG,ICLS,NCLS,NSBT,NELM,NRNGE1,NRNGE2,SGRMS,NICLS;
69000 ARRAY CLSMEM[*];
70000 BEGIN
71000    REAL NOSP,I,J,NSPECT,CPDLBL,NOPE,BWT,MWT,SC,SH,SO,SN,SS,SX,BC,BH,BO,BN,B
72000    S,BX,MS,INTMAX,ELM,K ;
73000    ARRAY CMMNT[0:13],CLS[1:7],MASS,INT[0:100];
74000    LABEL L1,L2 ;
75000    WRITE (LP,<//>) ;
76000    READ(TAPEDATA,<I3>,NOSP);
77000
78000    THRU NOSP DO
79000    BEGIN
80000       DO VECTORMODE (MASS,INT,FOR 101)
81000       BEGIN MASS:=INT:=0 ; INCREMENT MASS,INT ; END ;
82000       READ(TAPEDATA,<I3,A5,13A6,A2,3I3,7I2,/,12I3,/,(30F6.2)>,
83000       NSPECT,CPDLBL,CMMNT,NOPE,BWT,MWT,CLS,SC,SH,SO,SN,SS,SX,BC,BH,BO,BN,
84000       BS,BX,FOR I:=1 STEP 1 UNTIL NOPE DO [MASS[I],INT[I]]);
85000
86000       % CUTOFF BELOW M/Z 100,NORMALISE AND INT TSHLD=1%
87000       INTMAX:=0;
88000       FOR J:=1 STEP 1 UNTIL NOPE DO
89000       IF MASS[J]<100 THEN BEGIN NOPE:=J-1 ; GO TO L1 ; END
90000       ELSE
91000       INTMAX:=MAX(INTMAX,INT[J]);
92000       L1:;
93000       IF INTMAX^=100 THEN
94000       BEGIN
95000          FOR J:=1 STEP 1 UNTIL NOPE DO
96000          BEGIN INT[J]:=INT[J]*100/INTMAX;IF INT[J]<1 THEN INT[J]:=0;END;
97000          WRITE(LP,<I3," NORM FROM",I5>,NSPECT,INTMAX);
98000       END;
99000       %82 MOST POP M/Z
100000      I:=0;
101000      FOR K:=1 STEP 1 UNTIL 82 DO
102000      BEGIN
103000         I:=I+1;
104000         FOR J:=1 STEP 1 UNTIL NOPE DO IF MASS[J]=MZ82[K] THEN
105000         BEGIN DATA[NSPECT,I]:=INT[J];GO TO L2;END;
106000         L2:;
107000      END;
108000
109000      % CLASS ASSIGNMENT
110000      CLASS[NSPECT]:=-1;
111000      IF IBCL>0 THEN IF CLS[NCLS]>0 THEN CLASS[NSPECT]:=+1;
112000      IF IELM>0 THEN
113000      BEGIN
114000         CASE NSBT-1 OF
115000         BEGIN
116000            CASE NELM-1 OF
117000            BEGIN ELM:=SC;ELM:=SH;ELM:=SO;ELM:=SN;ELM:=SS;ELM:=SX;END;
118000            CASE NELM-1 OF
119000            BEGIN ELM:=BC;ELM:=BH;ELM:=BO;ELM:=BN;ELM:=BS;ELM:=BX;END;
120000            CASE NELM-1 OF
121000            BEGIN
122000               ELM:=BC+SC;ELM:=BH+SH;ELM:=BO+SO;ELM:=BN+SN;ELM:=BS+SS;ELM:=BX+SX;
123000            END;
124000         END;
125000         IF ELM>= NRNGE1 AND ELM<= NRNGE2 THEN CLASS[NSPECT]:=+1 ;
126000      END;
127000      IF ISUG>0 THEN IF MWT-BWT=SGRMS THEN CLASS[NSPECT]:=+1;
128000      IF ICLS>0 THEN
129000      FOR J:=1 STEP 1 UNTIL NICLS DO
130000      IF NSPECT=CLSMEM[J] THEN CLASS[NSPECT]:=+1 ;
131000   END;
132000 END ;
133000
```

```
134000 PROCEDURE ARRANG(IPRSET,NMEMTR,NNONTR,NMEM20,NNON20,NMEM49,NNON49);
135000
136000 %      ORDER AND COUNT FILE OF 125 SPECT
137000
138000 %      1-76 (TR76) MEM (NMEMTR) THEN NON MEM (NNONTR)
139000 %      77-96 (PR20) MEM (NMEM20) THEN NON MEM (NNON20)
140000 %      97-125 MEM THEN NON MEM
141000 %      NB. PR49 IS 77-125 AND CONTAINS NMEM49 MEM AND NNON49 NON MEM
142000 ARRAY IPRSET[*];REAL NMEMTR,NNONTR,NMEM20,NNON20,NMEM49,NNON49;
143000 BEGIN
144000    ARRAY ORIGINDEX[0:126];LABEL L1; REAL I,J ;
145000

146000    PROCEDURE SWOP(I1,I2);
147000    %      SWOP POSITIONS OF 2 VECTORS IN FILE
148000    REAL I1,I2;
149000    BEGIN
150000       ARRAY TEMPARAY,A,B[0:83] ; REAL TEMP ;
151000       TEMP:=CLASS[I2];
152000       DO VECTORMODE(TEMPARAY,DATA[I2,*],FOR 84)
153000       BEGIN TEMPARAY:=DATA;INCREMENT TEMPARAY,DATA;END;
154000       CLASS[I2]:=CLASS[I1];
155000       FOR J:=0 STEP 1 UNTIL 83 DO DATA[I2,J]:=DATA[I1,J] ;
156000       CLASS[I1]:=TEMP;
157000       FOR J:=0 STEP 1 UNTIL 83 DO DATA[I1,J]:=TEMPARAY[J] ;
158000       TEMP:=ORIGINDEX[I1] ;
159000       ORIGINDEX[I1]:=ORIGINDEX[I2] ;
160000       ORIGINDEX[I2]:=TEMP ;
161000    END;

162000    PROCEDURE MEMFIRST(I1,IMID,I2) ;
163000    %      ORDER MEMBERS FIRST IN TR SET
164000    REAL I1,IMID,I2 ;
165000    BEGIN
166000       REAL I,K ; LABEL L1,L2 ;
167000       FOR I:=I1 STEP 1 UNTIL IMID DO  IF CLASS[I]=-1 THEN
168000       BEGIN
169000          FOR K:=IMID+1 STEP 1 UNTIL I2 DO
170000          IF CLASS[K]=+1 THEN BEGIN SWOP(I,K) ; GO TO L1 ; END ;
171000          WRITE (LP,<//,"*** ERROR AT MEMFIRST OF ARRANG FOR SET",I4," -",I4,
172000          " ***",//>,  I1,I2) ;
173000          GO TO L2 ;
174000          L1 : ;
175000       END ;
176000       L2 : ;
177000    END ;
178000    FOR I:=1 STEP 1 UNTIL 125 DO ORIGINDEX[I]:=I ;
179000
180000    % TR 1-76 THEN PRED 77-125
181000    FOR I:=1 STEP 1 UNTIL 20 DO
182000    IF IPRSET[I]=76+I THEN GO TO L1 ELSE SWOP(IPRSET[I],76+I);
183000    L1:;
184000
185000    % COUNT MEM AND NON MEM
186000    NMEMTR:=0;
187000    FOR I:=1 STEP 1 UNTIL 76 DO IF CLASS[I]=+1 THEN NMEMTR:=*+1;
188000    NNONTR:=76-NMEMTR;
189000    NMEM20:=0;
190000    FOR I:=77 STEP 1 UNTIL 96 DO IF CLASS[I]=+1 THEN NMEM20:=*+1;
191000    NNON20:=20-NMEM20;
192000    NMEM49:=NMEM20;
193000    FOR I:=97 STEP 1 UNTIL 125 DO IF CLASS[I]=+1 THEN NMEM49:=*+1;
194000    NNON49:=49-NMEM49 ;
195000    WRITE (LP,<//,"GROUP SIZES",//,"TR",X6,"PR20",X4,"PR49",/,
196000    3("MEM NON "),//,6(I3,X1)>,NMEMTR,NNONTR,NMEM20,NNON20,NMEM49,NNON49);
197000
198000    % TR SET MEM FIRST
199000    % PR SET IN 2 PARTS 20 AND 49,MEM FIRST IN EACH
200000    MEMFIRST(1,NMEMTR,76) ;
201000    MEMFIRST(77,76+NMEM20,96) ;
202000    MEMFIRST(97,96+NMEM49-NMEM20,125) ;
203000
```

```
204000     %      OUTPUT MEM AND NON MEM IN EACH SET
205000     WRITE(LP,<//,"ORIGINAL SPECT NOS IN BRACKETS",//,"TR MEM",/,10(I5," (",I
206000     3,")")>,
207000     FOR I:=1 STEP 1 UNTIL NMEMTR DO [I,ORIGINDEX[I]]);
208000     WRITE(LP,<//,"TR NON MEM",/,10(I5," (",I3,")")>,
209000     FOR I:=1 STEP 1 UNTIL NNONTR DO [I,ORIGINDEX[NMEMTR+I]]);
210000     WRITE(LP,<//,"PR20 MEM",/,10(I5," (",I3,")")>,
211000     FOR I:=1 STEP 1 UNTIL NMEM20 DO[I,ORIGINDEX[76+I]]);
212000     WRITE(LP,<//,"PR20 NON MEM",/,10(I5," (",I3,")")>,
213000     FOR I:=1 STEP 1 UNTIL NNON20 DO [I,ORIGINDEX[76+NMEM20+I]]);
214000     WRITE(LP,<//,"PR49 2ND HALF MEM",/,10(I5," (",I3,")")>,
215000     FOR I:=1 STEP 1 UNTIL NMEM49-NMEM20 DO[I,ORIGINDEX[96+I]]);
216000     WRITE(LP,<//,"PR49 2ND HALF NON MEM",/,10(I5," (",I3,")")>,
217000     FOR I:=1 STEP 1 UNTIL NNON49-NNON20 DO [I,ORIGINDEX[96+NMEM49-NMEM20+I]]
218000     );
219000 END;
220000

221000 PROCEDURE AUTOSCALE(I1,I2,MEAN,STDDV);
222000
223000 % CENTRE AND NORM TR AND PR SETS
224000
225000 REAL I1,I2; ARRAY MEAN,STDDV[*];
226000 BEGIN
227000     REAL I,J;
228000     FOR J:=I1 STEP 1 UNTIL I2 DO
229000     FOR I:=1 STEP 1 UNTIL 82 DO MEAN[I]:=*+DATA[J,I];
230000     FOR I:=1 STEP 1 UNTIL 82 DO MEAN[I]:=MEAN[I]/(I2-I1+1);
231000     FOR J:=I1 STEP 1 UNTIL I2 DO
232000     FOR I:=1 STEP 1 UNTIL 82 DO STDDV[I]:=*+(DATA[J,I]-MEAN[I])**2;
233000     FOR I:=1 STEP 1 UNTIL 82 DO STDDV[I]:=SQRT(STDDV[I]/(I2-I1));
234000     FOR J:=I1 STEP 1 UNTIL I2 DO
235000     FOR I:=1 STEP 1 UNTIL 82 DO IF STDDV[I]>0 THEN
236000     DATA[J,I]:=(DATA[J,I]-MEAN[I])/STDDV[I];
237000     WRITE (LP,<//,"MEANS    ",//,(10F10.6)>,
238000     FOR I:=1 STEP 1 UNTIL 82 DO MEAN [I]) ;
239000     WRITE (LP,<//,"STD DEVS",//,(10F10.6)>,
240000     FOR I:=1 STEP 1 UNTIL 82 DO STDDV[I]) ;
241000 END;
242000

243000 PROCEDURE REGEN(I1,I2,MEAN,STDDV);
244000
245000 % REGENERATE DATA AND CONVERT TO BINARY
246000
247000 REAL I1,I2; ARRAY MEAN,STDDV[*];
248000 BEGIN
249000     REAL I,J;
250000     FOR J:=I1 STEP 1 UNTIL I2 DO
251000     FOR I:=1 STEP 1 UNTIL 82 DO
252000     IF DATA[J,I]:=DATA[J,I]*STDDV[I]+MEAN[I]>0 THEN
253000     DATA[J,I]:=1 ELSE DATA[J,I]:=0;
254000     FOR J:=I1 STEP 1 UNTIL I2 DO
255000     FOR I:=1 STEP 1  UNTIL 82 DO MEAN[I]:=*+DATA[J,I] ;
256000     FOR I:=1 STEP 1  UNTIL 82 DO MEAN[I]:=MEAN[I]/(I2-I1+1) ;
257000     WRITE (LP,<//,"BINARY MEANS",//,(10F10.6)>,
258000     FOR I:=1 STEP 1 UNTIL 82 DO MEAN[I]) ;
259000 END;
260000

261000 PROCEDURE EXEC(NMEMTR,NNONTR,NMEM20,NNON20,NMEM49,NNON49,NUM24,NUM12,
262000 AMU24,AMU12) ;
263000
264000 %      KNN AND DIST FROM MEAN CLASSIFS FOR VARIOUS M/Z SELECTIONS
265000 %      BIN AND LOG DATA
266000
267000 ARRAY AMU24,AMU12[*] ;
268000 REAL NMEMTR,NNONTR,NMEM20,NNON20,NMEM49,NNON49,NUM24,NUM12 ;
269000 BEGIN
270000     ARRAY MEANTR,MEAN49,STDDVTR,STDDV49,MZ[0:83] ;
271000     REAL I,J,K,NOMZ ;
272000
```

```
273000     PROCEDURE FEATSELECT(TOL) ;
274000     %      TAKE FULL 82, BEST 24 AND BEST 12 M/Z POSNS
275000     REAL TOL ;
276000     BEGIN
277000        NOMZ:=82 ;
278000        FOR I:=1 STEP 1 UNTIL 82 DO MZ[I]:=I ;
279000        DISTMEAN(NMEMTR,NNONTR,NMEM20,NNON20,NMEM49,NNON49,NOMZ,MZ,TOL ) ;
280000        NOMZ:=NUM24 ;
281000        K:=0 ;
282000        FOR I:=1 STEP 1 UNTIL 82 DO
283000        FOR J:=1 STEP 1  UNTIL NOMZ DO
284000        IF AMU24[J]=MZ82[I] THEN MZ[K:=*+1]:=I ;
285000        DISTMEAN(NMEMTR,NNONTR,NMEM20,NNON20,NMEM49,NNON49,NOMZ,MZ,TOL ) ;
286000        NOMZ:=NUM12 ;
287000        K:=0 ;
288000        FOR I:=1 STEP 1 UNTIL 82 DO
289000        FOR J:=1 STEP 1 UNTIL NOMZ DO
290000        IF AMU12[J]=MZ82[I] THEN MZ[K:=*+1]:=I ;
291000        DISTMEAN(NMEMTR,NNONTR,NMEM20,NNON20,NMEM49,NNON49,NOMZ,MZ,TOL ) ;
292000     END ;
293000
294000     %      AUTOSCALED LOG DATA
295000     WRITE (LP,<//,"ORIGINAL DATA">) ;
296000     WRITE (LP,<//,"TR76">) ;
297000     AUTOSCALE (1,76,MEANTR,STDDVTR) ;
298000     WRITE (LP,<//,"PR49">) ;
299000     AUTOSCALE (77,125,MEAN49,STDDV49) ;
300000     WRITE (LP,<//,125("*")>) ;
301000     WRITE (LP,<//,"AUTOSCALED LOG DATA">) ;
302000     FEATSELECT(0.2) ;
303000     %      BINARY DATA
304000     WRITE (LP,<//,125("*")>) ;
305000     WRITE (LP,<//,"BINARY SPECTRA">) ;
306000     WRITE (LP,<//,125("*")>) ;
307000     WRITE (LP,<//,"TR76">);
308000     REGEN (1,76,MEANTR,STDDVTR) ;
309000     WRITE (LP,<//,"PR49">);
310000     REGEN (77,125,MEAN49,STDDV49) ;
311000     FEATSELECT(0.05) ;
312000  END ;
313000

314000  PROCEDURE DISTMEAN(NMEMTR,NNONTR,NMEM20,NNON20,NMEM49,NNON49,NOMZ,MZ,
315000  TOL) ;
316000
317000  %      DISTANCE FROM MEAN AND KNN CLASSIFS
318000
319000  ARRAY MZ[*] ; REAL NMEMTR,NNONTR,NMEM20,NNON20,NMEM49,NNON49,NOMZ,TOL ;
320000  BEGIN
321000     ARRAY MEANMEM,MEANNON[0:84], ASSIGN,DISTMEM,DISTNON[1:125] ;
322000     ARRAY ASSDDZ[1:125], MZR[0:83], SETSIZ,MEMSIZ[0:2] ;
323000     REAL TRM1D,TRM2D,TRN1D,TRN2D,P20M1D,P20M2D,P20N1D,P20N2D,P49M1D,P49M2D,
324000     P49N1D,P49N2D,DDZ ;
325000     REAL I,J,TRM1,TRM2,TRN1,TRN2,P20M1,P20M2,P20N1,P20N2,P49M1,P49M2,
326000     P49N1,P49N2,NOMZR ;
327000     REAL T1,T2 ;
328000

329000     PROCEDURE OPCL(I1,I2,ADM,ADN,ADMD,ADND) ;
330000     %      OUTPUT RESULTS AND ADD SPECT CLASSIF IN EACH GROUP
331000     REAL I1,I2,ADM,ADN,ADMD,ADND ;
332000     BEGIN
333000        ADM:=ADN:=0 ;
334000        ADMD:=ADND:=0 ;
335000        FOR J:=I1 STEP 1 UNTIL I2 DO
336000        BEGIN
337000           WRITE (LP,<I3,2F11.6,2I5>,J,DISTMEM[J],DISTNON[J],ASSIGN[J],ASSDDZ[J]) ;
338000           IF ASSIGN[J]=1 THEN ADM:=*+1 ELSE ADN:=*+1 ;
339000           IF ASSDDZ[J]=+1 THEN ADMD:=*+1 ELSE IF ASSDDZ[J]=-1 THEN ADND:=*+1 ;
340000        END ;
341000        WRITE (LP,</>) ;
342000     END ;
343000
```

```
344000     PROCEDURE MEANTR(I1,I2,MEANS) ;
345000     %     MEANS OF MEM AND NON MEM OF TR SET
346000     REAL I1,I2 ; ARRAY MEANS[*] ;
347000     BEGIN
348000        FOR J:=I1 STEP 1 UNTIL I2 DO
349000        FOR I:=1 STEP 1 UNTIL NOMZ DO MEANS[MZ[I]]:=*+DATA[J,MZ[I]] ;
350000        FOR I:=1 STEP 1 UNTIL NOMZ DO MEANS[MZ[I]]:=MEANS[MZ[I]]/(I2-I1+1) ;
351000     END ;
352000
353000     PROCEDURE DISTANCE(NOMZZ,MZZ) ;
354000     %     CLASSIF BY DIST FROM MEANS
355000     REAL NOMZZ ; ARRAY MZZ[*] ;
356000     BEGIN
357000        IF TOL >=0.2 THEN
358000        BEGIN
359000           IF NOMZZ>50 THEN DDZ:=0.3 ELSE
360000           IF NOMZZ>15 THEN DDZ:=0.2 ELSE DDZ:=0.1 ;
361000        END
362000        ELSE
363000        BEGIN
364000           IF NOMZZ>50 THEN DDZ:=0.15 ELSE
365000           IF NOMZZ>15 THEN DDZ:=0.1 ELSE DDZ:=0.05 ;
366000        END ;
367000        FOR J:=1 STEP 1 UNTIL 125 DO
368000        BEGIN
369000           FOR I:=1 STEP 1 UNTIL NOMZZ DO
370000           BEGIN
371000              DISTMEM[J]:=*+(DATA[J,MZZ[I]]-MEANMEM[MZZ[I]])**2 ;
372000              DISTNON[J]:=*+(DATA[J,MZZ[I]]-MEANNON[MZZ[I]])**2 ;
373000           END ;
374000           DISTMEM[J]:=SQRT(DISTMEM[J]) ;
375000           DISTNON[J]:=SQRT(DISTNON[J]) ;
376000           IF DISTMEM[J]<DISTNON[J] THEN ASSIGN[J]:=+1 ELSE ASSIGN[J]:=-1 ;
377000           IF DISTMEM[J]-DISTNON[J]<-DDZ THEN ASSDDZ[J]:=+1 ELSE
378000           IF DISTMEM[J]-DISTNON[J]>+DDZ THEN ASSDDZ[J]:=-1 ELSE ASSDDZ[J]:=0 ;
379000        END ;
380000        %     OUTPUT RESULTS
381000        %     TRM1=NO. OF CLASS MEM OF TRSET CLASSIF AS SUCH
382000        %     TRM2=NO. OF CLASS MEM OF TRSET CLASSIF AS NON MEM
383000        %        ETC
384000        WRITE (LP,</,X4,
385000        2("DIST FROM  "), "CLASS",/, X8,"MEM",X7,"NON MEM",X4,"AS  DDZ",/>) ;
386000        WRITE (LP,</,"TR76",/>) ;
387000        OPCL(1,NMEMTR,TRM1,TRM2,TRM1D,TRM2D) ;
388000        OPCL(NMEMTR+1,76,TRN1,TRN2,TRN1D,TRN2D) ;
389000        WRITE (LP,</,"PR20",/>) ;
390000        OPCL(77,76+NMEM20,P20M1,P20M2,P20M1D,P20M2D) ;
391000        OPCL(77+NMEM20,96,P20N1,P20N2,P20N1D,P20N2D) ;
392000        WRITE (LP,</,"PR49 2ND HALF",/>) ;
393000        OPCL(97,96+NMEM49-NMEM20,P49M1,P49M2,P49M1D,P49M2D) ;
394000        OPCL(97+NMEM49-NMEM20,125,P49N1,P49N2,P49N1D,P49N2D) ;
395000
396000        P49M1:=*+P20M1 ; P49M2:=*+P20M2 ; P49N1:=*+P20N1 ; P49N2:=*+P20N2 ;
397000        P49M1D:=*+P20M1D ; P49N1D:=*+P20N1D ;
398000        P49M2D:=*+P20M2D ; P49N2D:=*+P20N2D ;
399000        WRITE (LP,<//,X19,"TR76", X15,"PR20",X15,"PR49",/,X15,
400000        3("MEM",X6,"NON",X7), /,"CLASSIF AS:",X2,3(2("MEM NON  "),X1), //,
401000        X13,3(2(I3,X1,I3,X2),X1)>,
402000        TRM1,TRM2,TRN1,TRN2,P20M1,P20M2,P20N1,P20N2,P49M1,P49M2,P49N1,P49N2) ;
403000        WRITE (LP,</,"DDZ",F5.2,X5,3(2(I3,X1,I3,X2),X1),/>,DDZ,
404000        TRM1D,TRM2D,TRN1D,TRN2D,P20M1D,P20M2D,P20N1D,P20N2D,P49M1D,P49M2D,
405000        P49N1D,P49N2D) ;
406000
407000        %     OUTPUT CLASSIF SUCCESS ON PUNCHED CARDS
408000        FILL SETSIZ WITH 76,20,49 ;
409000        WRITE (CP,<A6,12I3,X10,"DISTMEAN NO DDZ">,
410000        CATLBL,SETSIZ[0],NMEMTR,TRM2,TRN1,SETSIZ[1],NMEM20,P20M2,P20N1,
411000        SETSIZ[2],NMEM49,P49M2,P49N1) ;
412000        MEMSIZ[0]:=TRM1D+TRM2D ;
413000        MEMSIZ[1]:=P20M1D+P20M2D ;
414000        MEMSIZ[2]:=P49M1D+P49M2D ;
415000        SETSIZ[0]:=MEMSIZ[0]+TRN1D+TRN2D ;
416000        SETSIZ[1]:=MEMSIZ[1]+P20N1D+P20N2D ;
417000        SETSIZ[2]:=MEMSIZ[2]+P49N1D+P49N2D ;
418000        WRITE (CP,<A6,12I3,X10,"DISTMEAN DDZ",F5.2>,
419000        CATLBL,SETSIZ[0],MEMSIZ[0],TRM2D,TRN1D,SETSIZ[1],MEMSIZ[1],P20M2D,P20N1D
420000        ,SETSIZ[2],MEMSIZ[2],P49M2D,P49N1D,DDZ) ;
421000     END ;
422000
```

```
423000     PROCEDURE KNNPART(NOMZZ,MZZ) ;
424000     %      KNN CLASSIF FOR FULL AND REDUCED FEATURE SPACES
425000     REAL NOMZZ ; ARRAY MZZ[*] ;
426000     BEGIN
427000        ARRAY MMISCTR,MMISC20,MMISC49,NMISCTR,NMISC20,NMISC49[1:12] ;
428000        FILL SETSIZ WITH 76,20,49 ;
429000        %       TR SET RECOGNITION
430000        WRITE (LP,<//,"TR76">) ;
431000        T1:=TIME(2) ;
432000        KNN (1,NMEMTR,76,1,76,NOMZZ,MZZ,MMISCTR,NMISCTR) ;
433000        T2:=(TIME(2)-T1)/60 ; WRITE (LP,<"KNN TIME (SEC)",I5>,T2) ;
434000        %      PR20
435000        WRITE (LP,<//,"PR20">) ;
436000        T1:=TIME(2) ;
437000        KNN (77,76+NMEM20,96,1,76,NOMZZ,MZZ,MMISC20,NMISC20 ) ;
438000        T2:=(TIME(2)-T1)/60 ; WRITE (LP,<"KNN TIME (SEC)",I5>,T2) ;
439000        %      PR49 2ND HALF
440000        WRITE (LP,<//,"PR49 2ND HALF">) ;
441000        T1:=TIME(2) ;
442000        KNN (97,96+NMEM49-NMEM20,125,1,76,NOMZZ,MZZ,MMISC49,NMISC49) ;
443000        T2:=(TIME(2)-T1)/60 ; WRITE (LP,<"KNN TIME (SEC)",I5>,T2) ;
444000
445000        %      OUTPUT CLASSIF SUCCESS ON PUNCHED CARDS
446000        FOR I:=1 STEP 1 UNTIL 12 DO
447000        BEGIN MMISC49[I]:=*+MMISC20[I] ; NMISC49[I]:=*+NMISC20[I] ; END ;
448000        FOR I:=1 STEP 1 UNTIL 12 DO
449000        WRITE (CP,<A6,12I3,X10,"KNN NOMZ",I4>,
450000        CATLBL,SETSIZ[0],NMEMTR,MMISCTR[I],NMISCTR[I],SETSIZ[1],NMEM20,
451000        MMISC20[I],NMISC20[I],SETSIZ[2],NMEM49,MMISC49[I],NMISC49[I],NOMZZ) ;
452000     END ;
453000
454000     %      MEAN TR MEM AND NON MEM
455000     MEANTR(1,NMEMTR,MEANMEM) ;
456000     MEANTR(NMEMTR+1,76,MEANNON) ;
457000
458000     %      FORM REDUCED ARRAY (MZR) OF MZ82 INDICES
459000     %      EXCLUDING COMPTS OF MEAN NEARLY EQUAL IN MEM AND NON MEM
460000     FOR I:=1 STEP 1 UNTIL NOMZ DO MZR[I]:=MZ[I] ;
461000     NOMZR:=NOMZ ;
462000     FOR I:=1 STEP 1 UNTIL NOMZ DO
463000     IF MEANMEM[MZ[I]]>MEANNON[MZ[I]]-TOL AND
464000     MEANMEM[MZ[I]]<MEANNON[MZ[I]]+TOL THEN
465000     BEGIN
466000        NOMZR:=*-1 ;
467000        FOR J:=I STEP 1 UNTIL NOMZR DO  MZ[J]:=MZ[J+1] ;
468000     END;
469000
470000     %      DISTANCE FROM MEANS FOR ALL SPECT
471000     WRITE (LP,<//,125("*")>) ;
472000     WRITE (LP,<//,"DISTANCE FROM MEAN CLASSIF">) ;
473000     WRITE (LP,<//,"REDUCED MZ WITH EQUAL MEAN COMPTS REMOVED",//,"NOMZR",I3>
474000     , NOMZR) ;
475000     WRITE (LP,<//,"MZR",/,(30I4)>,
476000     FOR I:=1 STEP 1 UNTIL NOMZR DO MZ82[MZR[I]]) ;
477000     T1:=TIME(2) ;
478000     DISTANCE(NOMZR,MZR) ;
479000     T2:=(TIME(2)-T1)/60 ; WRITE (LP,<"DISTANCE TIME",I5>,T2) ;
480000
481000     WRITE (LP,<//,125("*")>) ;
482000     WRITE (LP,<//,"K NEAREST NEIGHBOUR CLASSIFICATION">) ;
483000     WRITE (LP,<//,"FULL SET OF MZ",//,"NOMZ",I3>,NOMZ) ;
484000     WRITE (LP,<//,"MZ",/,(30I4)>,
485000     FOR I:=1 STEP 1 UNTIL NOMZ  DO MZ82[MZ[I]]) ;
486000     KNNPART(NOMZ,MZ) ;
487000  END ;
488000
```

```
489000 PROCEDURE KNN(J1,JMID,J2,I1,I2,NOMZ,MZ,MEMMISC,NONMISC) ;
490000
491000 %     K NEAREST NEIGHBOUR CLASSIF K=1,3,5,7
492000
493000 REAL J1,JMID,J2,I1,I2,NOMZ ; ARRAY MZ,MEMMISC,NONMISC[*] ;
494000 BEGIN
495000    REAL I,J,K,L,FLG ;
496000    REAL T1,T2 ;
497000    ARRAY MEM,NON,ASN[0:4,1:3], DIST[0:76,0:76], MINDIST,MININDX[0:7] ;
498000    WRITE (LP, <//,X16,"K=1",X23,"K=3",X23,"K=5",X23,"K=7">) ;
499000    WRITE (LP,</,X8,4("SUM V SUM V/D SUMV/D2",X7),/>) ;
500000
501000    T1:=TIME(2) ;
502000    %     DISTANCE MATRIX BETWEEN VECTORS
503000    FOR J:=J1 STEP 1 UNTIL J2 DO  FOR I:=J-J1+1  STEP 1 UNTIL I2 DO
504000    BEGIN
505000       FOR K:=1 STEP 1 UNTIL NOMZ DO
506000       DIST[J-J1+1,I]:=*+(DATA[J,MZ[K]]-DATA[I,MZ[K]])**2 ;
507000       DIST[I,J-J1+1]:=DIST[J-J1+1,I]:=SQRT(DIST[J-J1+1,I]) ;
508000    END ;
509000    T2:=(TIME(2)-T1)/60 ; WRITE (LP,<"DIST MAT (SEC)",I5>,T2) ;
510000
511000    FOR J:=J1 STEP 1 UNTIL J2 DO
512000    BEGIN
513000
514000       %     1ST,2ND ... 7TH NEAREST NEIGHBOURS FOR EACH SPECT
515000       FILL MINDIST WITH 8(100) ; FILL MININDX WITH 8(0) ;
516000       FOR I:=I1 STEP 1 UNTIL I2 DO  IF DIST[J-J1+1,I]<MINDIST[1] THEN
517000       BEGIN MINDIST[1]:=DIST[J-J1+1,I] ; MININDX[1]:=I ; END ;
518000       FOR K:=2,3,4,5,6,7   DO FOR I:=I1 STEP 1 UNTIL I2 DO
519000       BEGIN
520000          FLG:=0 ;
521000          FOR L:=1 STEP 1 UNTIL K-1 DO  IF MININDX[L]=I THEN FLG:=1 ;
522000          IF FLG=0 THEN   IF DIST[J-J1+1,I]<MINDIST[K] THEN
523000          BEGIN MINDIST[K]:=DIST[J-J1+1,I] ; MININDX[K]:=I ; END ;
524000       END ;
525000
526000       %     SUMS OF VOTES AND WEIGHTED VOTES FOR EACH SPECT FOR K=1,3,5,7
527000       FOR K:=1,3,5,7 DO
528000       BEGIN
529000          FOR L:=1,2,3 DO ASN[(K+1)/2,L]:=ASN[(K-1)/2,L] ;
530000          FOR L:=K-1,K DO
531000          BEGIN
532000             ASN[(K+1)/2,1]:=*+CLASS[MININDX[L]] ;
533000             IF DIST[J-J1+1,MININDX[L]]>0 THEN
534000             BEGIN
535000                ASN[(K+1)/2,2]:=*+CLASS[MININDX[L]]/DIST[J-J1+1,MININDX[L]] ;
536000                ASN[(K+1)/2,3]:=*+CLASS[MININDX[L]]/DIST[J-J1+1,MININDX[L]]**2 ;
537000             END ;
538000          END ;
539000          FOR L:=1,2,3 DO  IF ASN[(K+1)/2,L]>0 THEN
540000          MEM[(K+1)/2,L]:=*+1 ELSE NON[(K+1)/2,L]:=*+1 ;
541000       END ;
542000
543000       %     OUTPUT RESULTS
544000       WRITE (LP,<I3,X2,4(3F8.4,X4)>,
545000       J, FOR K:=1,2,3,4 DO [ FOR L:=1,2,3 DO ASN[K,L] ] ) ;
546000       IF J=JMID OR J=J2 THEN
547000       BEGIN
548000          WRITE (LP,</,"MEM  ",4(3I8,X4),/>,FOR K:=1,2,3,4 DO MEM[K,*] ) ;
549000          WRITE (LP,</,"NON  ",4(3I8,X4),/>,FOR K:=1,2,3,4 DO NON[K,*] ) ;
550000          I:=0 ;
551000          IF J=JMID THEN
552000          FOR K:=1,2,3,4 DO FOR L:=1,2,3 DO MEMMISC[I:=*+1]:=NON[K,L]
553000          ELSE
554000          FOR K:=1,2,3,4 DO FOR L:=1,2,3 DO NONMISC[I:=*+1]:=MEM[K,L] ;
555000          FOR K:=1,2,3,4 DO
556000          BEGIN FILL MEM[K,*] WITH 3(0) ; FILL NON[K,*] WITH 3(0) ; END ;
557000       END ;
558000    END ;
559000 END ;
```

```
O FUNCT AT C2

SHORT TITLE   OC2


INPUT PARAMETERS

  IBCL  IELM  ISUG  ICLS

     0     0     0     1

  NCLS  NSBT  NELM RNGE1 RNGE2 SGRMS NICLS NUM24 NUM12

     0     0     0     0     0     0    31    24     8

AMU24

 112 113 125 126 127 133 136 140 146 148 151 163 164 168 169 171 185 193 207 211 218 219 227 228

AMU12

 112 125 127 140 151 168 169 228

CLSMEM

   2   7  26  27  28  29  30  31  32  33  34  35  37  38  39  40  56  59  60  61  84  86  94  96 102 106 111 116 117 118
 124

PRSET

  61  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106
 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125
```

```
  1 NORM FROM   70
  2 NORM FROM   50
  5 NORM FROM   70
  7 NORM FROM   81
 17 NORM FROM   80
 26 NORM FROM   60
 28 NORM FROM   20
 30 NORM FROM   90
 34 NORM FROM   27
 35 NORM FROM   20
 40 NORM FROM   15
 50 NORM FROM   85
 54 NORM FROM   75
 55 NORM FROM   80
 56 NORM FROM   48
 59 NORM FROM   60
 61 NORM FROM   45
 68 NORM FROM   80
 81 NORM FROM   25
 82 NORM FROM   94
 84 NORM FROM   82
 85 NORM FROM   84
 88 NORM FROM   80
 95 NORM FROM   20
116 NORM FROM   90

TAPEINPUT TIME   11


GROUP SIZES

TR       PR20     PR49
MEM NON MEM NON MEM NON

 19  57   5  15  12  37


ORIGINAL SPECT NOS IN BRACKETS

TR MEM
    1 ( 26)    2 (  2)    3 ( 27)    4 ( 28)    5 ( 29)    6 ( 30)    7 (  7)    8 ( 31)    9 ( 32)   10 ( 33)
   11 ( 34)   12 ( 35)   13 ( 37)   14 ( 38)   15 ( 39)   16 ( 40)   17 ( 56)   18 ( 59)   19 ( 60)

TR NON MEM
    1 ( 20)    2 ( 21)    3 ( 22)    4 ( 23)    5 ( 24)    6 ( 25)    7 (  1)    8 (  3)    9 (  4)   10 (  5)
   11 (  6)   12 (  8)   13 (  9)   14 ( 10)   15 ( 11)   16 ( 12)   17 ( 36)   18 ( 13)   19 ( 14)   20 ( 15)
   21 ( 16)   22 ( 41)   23 ( 42)   24 ( 43)   25 ( 44)   26 ( 45)   27 ( 46)   28 ( 47)   29 ( 48)   30 ( 49)
   31 ( 50)   32 ( 51)   33 ( 52)   34 ( 53)   35 ( 54)   36 ( 55)   37 ( 17)   38 ( 57)   39 ( 58)   40 ( 18)
   41 ( 19)   42 ( 77)   43 ( 62)   44 ( 63)   45 ( 64)   46 ( 65)   47 ( 66)   48 ( 67)   49 ( 68)   50 ( 69)
   51 ( 70)   52 ( 71)   53 ( 72)   54 ( 73)   55 ( 74)   56 ( 75)   57 ( 76)
```

```
PR20 MEM
    1 ( 61)    2 ( 84)    3 ( 86)    4 ( 94)    5 ( 96)

PR20 NON MEM
    1 ( 82)    2 ( 83)    3 ( 78)    4 ( 85)    5 ( 79)    6 ( 87)    7 ( 88)    8 ( 89)    9 ( 90)   10 ( 91)
   11 ( 92)   12 ( 93)   13 ( 80)   14 ( 95)   15 ( 81)

PR49 2ND HALF MEM
    1 (106)    2 (111)    3 (116)    4 (117)    5 (118)    6 (102)    7 (124)

PR49 2ND HALF NON MEM
    1 (104)    2 (105)    3 ( 97)    4 (107)    5 (108)    6 (109)    7 (110)    8 ( 98)    9 (112)   10 (113)
   11 (114)   12 (115)   13 ( 99)   14 (100)   15 (101)   16 (119)   17 (120)   18 (121)   19 (122)   20 (123)
   21 (103)   22 (125)

ARRANG    TIME    0


ORIGINAL DATA


TR76

MEANS

  5.511435  0.921053  2.017544  0.907895  7.721126  4.039220  1.229360  0.885965  5.234023  3.755639
  1.835526  1.991486  2.229323  8.021930  6.951267  1.815789  9.421053 22.764024 15.438682  3.820175
  2.521930  2.075049  4.999472  8.867020  5.220809  5.750000  5.618421  1.948830  3.439327  2.151316
  4.270898  5.856811 18.177825  3.274768  2.131579  0.318713  1.412240  0.565789  0.626259  0.593358
  0.959586  8.170426  0.843567  0.960526  1.907895  1.007269  1.303627  1.736326  5.316434  1.653509
  1.495614  0.785088  1.916151  2.180626  3.742239  1.247563  0.865132  0.701917  2.273392  2.419799
  1.605263  5.119711  1.911184  3.497563  0.915367  0.228964  0.304987  2.983414  0.637752  3.960526
  0.171053  2.753096  1.094234  0.418546  0.913957  0.644737  1.146478  0.761278  0.184211  1.059598
  1.125387  0.131579

STD DEVS

  9.963948  4.374969 10.807253  4.179881 24.309033 15.687463  5.609110  4.380953 20.074774  9.990140
  5.468783 11.684245 12.190418 25.788733 21.985863  8.856953 27.381874 35.679464 30.325440 16.082401
  8.610646 11.716925 17.947437 21.370759 16.476517 21.619821 20.326644 11.735752 14.329480 11.156768
 16.517633 16.372116 33.602736 15.542874 12.019254  1.822885  4.752989  3.155252  2.562996  2.581900
  3.241154 20.749216  5.833412  5.234350 11.352741  3.437474  4.410402  6.384348 17.278965  8.448790
  5.921835  3.937175  8.319604  9.715164 11.939619  7.667413  3.884465  3.370404 12.209953 12.271969
  5.070382 17.991152 11.833547 16.574187  3.729680  0.855578  1.197191 12.464557  2.458144 16.944176
  0.806443 13.608048  3.625181  2.033555  4.218057  2.942579  5.860909  2.671124  0.811756  4.354504
  4.042078  0.680041
```

PR49

MEANS

5.479592	2.838037	2.531915	2.408163	5.970190	3.188153	1.000000	1.428571	4.102041	3.533218
4.510204	3.556122	6.077664	5.666667	4.802721	1.102041	2.438641	20.318281	12.244898	1.043084
4.591837	5.560091	1.698357	6.956687	12.908163	5.301129	2.824101	2.244898	0.829268	7.002605
1.469388	6.602041	8.239796	9.441550	4.102041	0.489796	3.149328	1.102041	1.997013	1.724490
1.244898	10.793651	6.462018	3.706904	0.428571	0.331010	2.632653	0.775510	2.349206	0.260204
3.558945	1.737468	0.668367	3.495175	4.173686	0.648285	5.517572	0.858446	2.061224	1.826531
2.367347	8.065112	1.515415	1.632653	0.666556	0.024888	0.331010	4.006513	0.927438	4.346613
1.671257	5.413808	6.967000	0.102041	0.160942	0.489796	0.122449	0.673469	0.510204	2.186279
1.000000	0.000000								

STD DEVS

8.761872	6.934911	9.429577	8.594859	20.915431	14.587691	3.378856	3.952847	19.987751	6.027419
9.674112	7.336990	20.688926	18.008100	16.390760	4.519610	8.630362	31.062954	27.397179	4.378317
19.252617	20.746733	6.065872	12.244650	30.742152	16.132499	9.982237	14.327320	2.973710	24.150888
3.857363	21.077481	22.364055	22.981194	14.606284	2.475904	15.263344	4.848732	8.487642	6.496041
4.815473	21.342627	21.635690	9.419459	2.000000	1.588490	8.161632	3.820260	6.759310	0.886765
12.347906	5.138230	2.192190	14.199985	14.093388	2.395205	18.343829	3.234006	6.835594	7.332541
6.290515	24.101213	5.633938	11.428571	3.072447	0.174216	1.250986	12.795457	3.809544	15.890693
8.771761	19.169495	17.720825	0.714286	0.964607	2.209095	0.389051	3.030317	3.571429	14.272144
3.034524	0.000000								

****

AUTOSCALED LOG DATA

****

DISTANCE FROM MEAN CLASSIF

REDUCED MZ WITH EQUAL MEAN COMPTS REMOVED

NOMZR 69

MZR

108 109 110 111 112 113 114 115 117 119 120 121 125 126 127 133 134 135 136 139 140 141 146 148 149 151 152 154 155 160
162 163 164 165 166 168 169 170 171 176 177 178 179 180 183 185 190 191 192 193 194 200 201 202 206 207 208 209 211 218
219 220 221 225 226 227 228 232 240

DIST FROM MEM	DIST FROM NON MEM	CLASS AS	DDZ

1	10.855837	11.882069	1	1
2	8.324484	7.828936	-1	-1
3	10.948571	12.346535	1	1
4	10.376348	12.038618	1	1
5	5.516194	6.742549	1	1
6	5.696649	6.151726	1	1
7	6.918929	8.539949	1	1
8	5.311784	5.931430	1	1
9	7.588571	8.987396	1	1
10	7.436817	8.292784	1	1
11	5.863029	6.214899	1	1
12	12.883909	14.318175	1	1
13	10.906449	12.043064	1	1
14	9.281300	10.139944	1	1
15	9.621042	10.145299	1	1
16	8.713452	9.873804	1	1
17	11.346933	12.681498	1	1
18	5.913870	6.064040	1	0
19	8.160573	8.553913	1	1
20	6.295082	4.875499	-1	-1
21	6.286283	4.886147	-1	-1
22	15.271050	14.061686	-1	-1
23	10.588705	8.864940	-1	-1
24	6.967069	5.267343	-1	-1
25	11.600470	11.054794	-1	-1
26	8.326341	7.452455	-1	-1
27	7.334338	5.967007	-1	-1
28	6.622579	5.209434	-1	-1
29	11.029290	9.465563	-1	-1
30	7.653190	6.980330	-1	-1
31	7.412202	5.036638	-1	-1
32	5.606211	3.501164	-1	-1
33	6.874638	4.541610	-1	-1
34	5.915918	3.964811	-1	-1
35	5.843146	3.607351	-1	-1
36	6.624338	6.149747	-1	-1
37	5.715323	3.392830	-1	-1
38	7.172355	5.225770	-1	-1
39	9.514992	8.136091	-1	-1
40	8.521914	6.901234	-1	-1
41	7.206722	4.835928	-1	-1
42	8.019167	6.146604	-1	-1
43	12.427994	11.476689	-1	-1
44	5.294625	3.379425	-1	-1
45	6.669577	6.695324	1	0
46	17.850426	18.717451	1	1
47	10.022572	9.057313	-1	-1
48	4.878889	4.041305	-1	-1
49	4.818499	3.265976	-1	-1
50	14.299613	13.195946	-1	-1

51	8.572284	7.255423	-1	-1
52	6.288160	4.055162	-1	-1
53	9.276111	8.301705	-1	-1
54	10.542223	9.613642	-1	-1
55	7.078820	5.329388	-1	-1
56	7.011478	5.489387	-1	-1
57	11.270530	10.121162	-1	-1
58	7.405561	6.012156	-1	-1
59	7.972070	6.495676	-1	-1
60	10.647964	9.406426	-1	-1
61	7.059102	5.896013	-1	-1
62	13.283609	12.082219	-1	-1
63	8.065140	7.871095	-1	0
64	8.854622	8.355073	-1	-1
65	10.326159	9.697274	-1	-1
66	9.033551	7.609474	-1	-1
67	9.206040	8.817404	-1	-1
68	6.361667	4.697876	-1	-1
69	10.213086	9.786084	-1	-1
70	5.139536	3.589038	-1	-1
71	9.543351	9.171703	-1	-1
72	11.409834	10.930173	-1	-1
73	4.055732	2.448976	-1	-1
74	6.245623	5.217953	-1	-1
75	8.100580	6.933214	-1	-1
76	8.100580	6.933214	-1	-1

PR20

77	7.876654	8.290328	1	1
78	9.439342	10.387442	1	1
79	5.972071	5.828319	-1	0
80	7.369434	8.280265	1	1
81	6.261449	6.768911	1	1
82	7.325493	7.004940	-1	-1
83	6.249310	4.853025	-1	-1
84	6.130490	4.948574	-1	-1
85	8.535983	7.486119	-1	-1
86	9.108707	8.373565	-1	-1
87	9.987451	9.169225	-1	-1
88	10.234474	8.958798	-1	-1
89	4.819643	4.255327	-1	-1
90	7.400452	7.000810	-1	-1
91	6.396573	5.001505	-1	-1
92	7.119353	5.194675	-1	-1
93	8.516838	7.751167	-1	-1
94	6.378537	4.920636	-1	-1
95	8.055414	7.006775	-1	-1
96	9.158091	8.820646	-1	-1

PR49 2ND HALF

97	10.278751	10.264426	-1	0
98	10.339853	10.822691	1	1
99	8.306594	9.669865	1	1
100	10.675501	11.696830	1	1
101	10.203021	10.711279	1	1
102	9.209798	9.633090	1	1
103	9.599249	9.948999	1	1
104	8.645882	7.246662	-1	-1
105	8.814525	7.451151	-1	-1
106	5.003848	3.991431	-1	-1
107	7.453497	6.677470	-1	-1
108	10.658764	10.136988	-1	-1
109	6.324974	5.175666	-1	-1
110	6.424425	5.301528	-1	-1
111	10.387378	9.775224	-1	-1
112	6.912680	6.612428	-1	-1
113	11.661348	9.766676	-1	-1
114	10.862465	9.483541	-1	-1
115	11.297907	9.988852	-1	-1
116	9.998076	9.932186	-1	0
117	8.102391	6.384307	-1	-1
118	8.463296	6.668356	-1	-1
119	6.775212	5.134122	-1	-1
120	12.365422	11.191755	-1	-1
121	8.534447	8.380425	-1	0
122	5.277121	4.808104	-1	-1
123	11.365174	10.950689	-1	-1
124	16.217687	15.972557	-1	0
125	4.614922	2.989625	-1	-1

	TR76				PR20				PR49			
	MEM		NON		MEM		NON		MEM		NON	
CLASSIF AS:	MEM	NON	MEM	NON	MEM	NON	MEM	NON	MEM	NON	MEM	NON
	18	1	2	55	4	1	0	15	10	2	0	37
DDZ 0.30	17	1	1	54	4	0	0	15	10	0	0	34

DISTANCE TIME 2

****

K NEAREST NEIGHBOUR CLASSIFICATION

FULL SET OF MZ

NOMZ 82

MZ
108 109 111 112 113 114 115 117 119 120 121 125 126 127 133 134 135 136 140 141 146 148 149 151 152 155 160 162 163 164
165 168 169 170 171 176 177 178 180 185 190 192 193 194 201 202 207 209 211 218 219 220 221 226 227 228 232 240 248 249
250 251 258 266 267 269 280 281 316 316 316 316 316 316 316 316 316 316 316 316 316 316

TR76

	K=1			K=3			K=5			K=7		
	SUM V	SUM V/D	SUMV/D2	SUM V	SUM V/D	SUMV/D2	SUM V	SUM V/D	SUMV/D2	SUM V	SUM V/D	SUMV/D2
DIST MAT (SEC)	28											
1	1.0000	0.0000	0.0000	1.0000	0.0077	0.0013	-1.0000	-0.1555	-0.0120	-1.0000	-0.1552	-0.0120
2	1.0000	0.0000	0.0000	-1.0000	-0.3246	-0.0536	-3.0000	-0.5798	-0.0862	-5.0000	-0.8266	-0.1166
3	1.0000	0.0000	0.0000	3.0000	0.1964	0.0193	1.0000	0.0297	0.0054	1.0000	0.0294	0.0053
4	1.0000	0.0000	0.0000	1.0000	-0.0033	-0.0006	1.0000	0.0015	0.0002	-1.0000	-0.1574	-0.0124
5	1.0000	0.0000	0.0000	3.0000	0.2797	0.0392	5.0000	0.5174	0.0675	3.0000	0.2867	0.0409
6	1.0000	0.0000	0.0000	1.0000	0.0878	0.0354	-1.0000	-0.2080	-0.0083	-3.0000	-0.4914	-0.0485
7	1.0000	0.0000	0.0000	3.0000	0.2288	0.0262	1.0000	0.0057	0.0013	-1.0000	-0.2101	-0.0220
8	1.0000	0.0000	0.0000	-1.0000	-0.3092	-0.0478	-3.0000	-0.6036	-0.0912	-3.0000	-0.6037	-0.0912
9	1.0000	0.0000	0.0000	3.0000	0.2448	0.0300	5.0000	0.4625	0.0537	3.0000	0.2539	0.0319
10	1.0000	0.0000	0.0000	1.0000	0.0028	0.0007	1.0000	0.0041	0.0010	1.0000	0.0025	0.0006
11	1.0000	0.0000	0.0000	-1.0000	-0.3498	-0.0612	-1.0000	-0.3476	-0.0605	-3.0000	-0.6658	-0.1111
12	1.0000	0.0000	0.0000	1.0000	0.0185	0.0034	3.0000	0.1807	0.0166	3.0000	0.1805	0.0166
13	1.0000	0.0000	0.0000	1.0000	-0.0025	-0.0004	1.0000	0.0003	0.0000	-1.0000	-0.1599	-0.0128
14	1.0000	0.0000	0.0000	-1.0000	-0.2174	-0.0240	-1.0000	-0.2176	-0.0240	-3.0000	-0.4037	-0.0414
15	1.0000	0.0000	0.0000	3.0000	0.2084	0.0217	3.0000	0.2092	0.0219	1.0000	0.0201	0.0040
16	1.0000	0.0000	0.0000	1.0000	-0.0014	-0.0003	-1.0000	-0.2044	-0.0209	-3.0000	-0.4058	-0.0412
17	1.0000	0.0000	0.0000	1.0000	0.0026	0.0004	1.0000	0.0031	0.0005	1.0000	0.0034	0.0006
18	1.0000	0.0000	0.0000	3.0000	0.4159	0.0893	1.0000	0.1046	0.0409	-1.0000	-0.1883	-0.0021
19	1.0000	0.0000	0.0000	-1.0000	-0.2621	-0.0343	-3.0000	-0.4941	-0.0613	-3.0000	-0.4957	-0.0616
MEM	19	0	0	14	11	11	11	11	11	7	7	7
NON	0	19	19	5	8	8	8	8	8	12	12	12

20	-1.0000	0.0000	0.0000	-3.0000 -6.9424-45.2150	-5.0000 -7.2905-45.2756	-7.0000 -7.6112-45.3271
21	-1.0000	0.0000	0.0000	-3.0000 -6.9382-45.2131	-5.0000 -7.2857-45.2735	-7.0000 -7.6069-45.3252
22	-1.0000	0.0000	0.0000	-3.0000 -0.1723 -0.0149	-5.0000 -0.3323 -0.0277	-7.0000 -0.4912 -0.0403
23	-1.0000	0.0000	0.0000	-3.0000 -0.1718 -0.0148	-5.0000 -0.3266 -0.0267	-7.0000 -0.4799 -0.0385
24	-1.0000	0.0000	0.0000	-3.0000 -0.2244 -0.0252	-5.0000 -0.4369 -0.0478	-7.0000 -0.6438 -0.0692
25	-1.0000	0.0000	0.0000	-3.0000 -0.2478 -0.0307	-5.0000 -0.4899 -0.0600	-7.0000 -0.7255 -0.0878
26	-1.0000	0.0000	0.0000	-3.0000 -0.2998 -0.0465	-5.0000 -0.5405 -0.0754	-7.0000 -0.7782 -0.1037
27	-1.0000	0.0000	0.0000	-3.0000 -0.4204 -0.0900	-5.0000 -0.6994 -0.1290	-7.0000 -0.9629 -0.1637
28	-1.0000	0.0000	0.0000	-3.0000 -0.4508 -0.1020	-5.0000 -0.8361 -0.1763	-7.0000 -1.1878 -0.2381
29	-1.0000	0.0000	0.0000	-3.0000 -0.1989 -0.0198	-5.0000 -0.3928 -0.0386	-7.0000 -0.5821 -0.0565
30	-1.0000	0.0000	0.0000	-3.0000 -0.4600 -0.1072	-5.0000 -0.8301 -0.1758	-7.0000 -1.1769 -0.2359
31	-1.0000	0.0000	0.0000	-3.0000 -0.3296 -0.0545	-5.0000 -0.6278 -0.0991	-7.0000 -0.9101 -0.1389
32	-1.0000	0.0000	0.0000	-3.0000 -0.5868 -0.1760	-5.0000 -1.0484 -0.2828	-7.0000 -1.4500 -0.3638
33	-1.0000	0.0000	0.0000	-3.0000 -0.5622 -0.1591	-5.0000 -1.0352 -0.2713	-7.0000 -1.4698 -0.3658
34	-1.0000	0.0000	0.0000	-3.0000 -0.5380 -0.1454	-5.0000 -0.9994 -0.2521	-7.0000 -1.4169 -0.3393
35	-1.0000	0.0000	0.0000	-3.0000 -0.3409 -0.0583	-5.0000 -0.6540 -0.1074	-7.0000 -0.9607 -0.1544
36	-1.0000	0.0000	0.0000	-1.0000 0.0117 0.0042	-3.0000 -0.3192 -0.0506	-5.0000 -0.6385 -0.1016
37	-1.0000	0.0000	0.0000	-3.0000 -0.4601 -0.1059	-5.0000 -0.9089 -0.2066	-7.0000 -1.3422 -0.3005
38	-1.0000	0.0000	0.0000	-3.0000 -0.3676 -0.0676	-5.0000 -0.7117 -0.1268	-7.0000 -1.0386 -0.1802
39	-1.0000	0.0000	0.0000	-3.0000 -0.3012 -0.0458	-5.0000 -0.5483 -0.0764	-7.0000 -0.7865 -0.1048
40	-1.0000	0.0000	0.0000	-3.0000 -0.4677 -0.1196	-5.0000 -0.7461 -0.1584	-7.0000 -1.0167 -0.1950
41	-1.0000	0.0000	0.0000	-3.0000 -0.5918 -0.1753	-5.0000 -1.0222 -0.2682	-7.0000 -1.3953 -0.3377
42	-1.0000	0.0000	0.0000	-3.0000 -0.2712 -0.0368	-5.0000 -0.5150 -0.0666	-7.0000 -0.7409 -0.0921
43	-1.0000	0.0000	0.0000	-3.0000 -0.2491 -0.0312	-5.0000 -0.4770 -0.0571	-7.0000 -0.6943 -0.0808
44	-1.0000	0.0000	0.0000	-3.0000 -0.4553 -0.1037	-5.0000 -0.8709 -0.1903	-7.0000 -1.2573 -0.2649
45	-1.0000	0.0000	0.0000	-3.0000 -0.3339 -0.0558	-5.0000 -0.6541 -0.1070	-7.0000 -0.9602 -0.1539
46	-1.0000	0.0000	0.0000	-1.0000 0.0020 0.0003	1.0000 0.1214 0.0074	3.0000 0.2394 0.0143
47	-1.0000	0.0000	0.0000	-3.0000 -0.2267 -0.0260	-5.0000 -0.4279 -0.0462	-5.0000 -0.4292 -0.0465
48	-1.0000	0.0000	0.0000	-3.0000 -0.1699 -0.0144	-5.0000 -0.3074 -0.0241	-5.0000 -0.3075 -0.0241
49	-1.0000	0.0000	0.0000	-3.0000 -0.4069 -0.0828	-5.0000 -0.8031 -0.1613	-5.0000 -0.8222 -0.1684
50	-1.0000	0.0000	0.0000	-3.0000 -0.1505 -0.0114	-5.0000 -0.2829 -0.0202	-7.0000 -0.4112 -0.0284
51	-1.0000	0.0000	0.0000	-3.0000 -0.2288 -0.0263	-5.0000 -0.4306 -0.0467	-7.0000 -0.6267 -0.0659
52	-1.0000	0.0000	0.0000	-3.0000 -0.5455 -0.1492	-5.0000 -1.0374 -0.2703	-7.0000 -1.4781 -0.3674
53	-1.0000	0.0000	0.0000	-3.0000 -0.2232 -0.0249	-5.0000 -0.4410 -0.0486	-7.0000 -0.6574 -0.0720
54	-1.0000	0.0000	0.0000	-3.0000 -0.2313 -0.0268	-3.0000 -0.2283 -0.0262	-5.0000 -0.4192 -0.0444
55	-1.0000	0.0000	0.0000	-3.0000 -0.3685 -0.0679	-5.0000 -0.6794 -0.1162	-7.0000 -0.9704 -0.1586
56	-1.0000	0.0000	0.0000	-3.0000 -0.5176 -0.1383	-5.0000 -0.8470 -0.1926	-7.0000 -1.1625 -0.2424
57	-1.0000	0.0000	0.0000	-3.0000 -0.1758 -0.0155	-5.0000 -0.3467 -0.0301	-7.0000 -0.5142 -0.0441
58	-1.0000	0.0000	0.0000	-3.0000 -0.3554 -0.0634	-5.0000 -0.6317 -0.1016	-7.0000 -0.8971 -0.1368
59	-1.0000	0.0000	0.0000	-3.0000 -0.2271 -0.0258	-3.0000 -0.2295 -0.0263	-5.0000 -0.4263 -0.0457
60	-1.0000	0.0000	0.0000	-3.0000 -0.2108 -0.0222	-5.0000 -0.4169 -0.0435	-7.0000 -0.6137 -0.0628
61	-1.0000	0.0000	0.0000	-3.0000 -0.3711 -0.0695	-5.0000 -0.7032 -0.1247	-7.0000 -0.9875 -0.1651
62	-1.0000	0.0000	0.0000	-3.0000 -0.1454 -0.0106	-5.0000 -0.2849 -0.0203	-7.0000 -0.4208 -0.0295
63	-1.0000	0.0000	0.0000	-3.0000 -0.3317 -0.0551	-5.0000 -0.6295 -0.0995	-5.0000 -0.6304 -0.0997
64	-1.0000	0.0000	0.0000	-3.0000 -0.3175 -0.0504	-5.0000 -0.5809 -0.0854	-7.0000 -0.8105 -0.1118
65	-1.0000	0.0000	0.0000	-3.0000 -0.4213 -0.0930	-5.0000 -0.7288 -0.1403	-5.0000 -0.7297 -0.1406
66	-1.0000	0.0000	0.0000	-3.0000 -0.2480 -0.0307	-5.0000 -0.4719 -0.0558	-7.0000 -0.6718 -0.0758
67	-1.0000	0.0000	0.0000	1.0000 0.3178 0.0505	-1.0000 0.0627 0.0179	-3.0000 -0.1760 -0.0106
68	-1.0000	0.0000	0.0000	-3.0000 -0.6244 -0.1962	-5.0000 -1.0807 -0.3006	-7.0000 -1.4527 -0.3698
69	-1.0000	0.0000	0.0000	-3.0000 -0.1994 -0.0199	-5.0000 -0.3718 -0.0348	-7.0000 -0.5371 -0.0485
70	-1.0000	0.0000	0.0000	-3.0000 -0.3011 -0.0454	-5.0000 -0.5745 -0.0828	-7.0000 -0.8412 -0.1184
71	-1.0000	0.0000	0.0000	-3.0000 -0.2385 -0.0289	-5.0000 -0.4444 -0.0501	-5.0000 -0.4451 -0.0502
72	-1.0000	0.0000	0.0000	-3.0000 -0.2533 -0.0321	-5.0000 -0.4957 -0.0615	-5.0000 -0.4962 -0.0616
73	-1.0000	0.0000	0.0000	-3.0000 -0.1317 -0.0095	-5.0000 -0.2119 -0.0128	-7.0000 -0.2839 -0.0154
74	-1.0000	0.0000	0.0000	-3.0000 -0.3457 -0.0598	-5.0000 -0.6720 -0.1130	-7.0000 -0.9911 -0.1639
75	-1.0000	0.0000	0.0000	-3.0000 -0.1661 -0.0276	-5.0000 -0.4323 -0.0644	-7.0000 -0.6375 -0.0854
76	-1.0000	0.0000	0.0000	-3.0000 -0.1661 -0.0276	-5.0000 -0.4323 -0.0644	-7.0000 -0.6375 -0.0854

MEM	0	0	0	1	3	3	1	2	2	1	1	1
NON	57	57	57	56	54	54	56	55	55	56	56	56

KNN TIME (SEC) 36

PR20

	K=1			K=3			K=5			K=7		
	SUM V	SUM V/D	SUMV/D2	SUM V	SUM V/D	SUMV/D2	SUM V	SUM V/D	SUMV/D2	SUM V	SUM V/D	SUMV/D2
DIST MAT (SEC)	12											
77	1.0000	0.1985	0.0394	-1.0000	-0.1030	-0.0061	-1.0000	-0.1059	-0.0069	-3.0000	-0.3848	-0.0458
78	1.0000	0.1264	0.0160	1.0000	0.1136	0.0131	1.0000	0.1137	0.0132	-1.0000	-0.0690	-0.0035
79	1.0000	0.4713	0.2222	1.0000	0.6080	0.2865	1.0000	0.6056	0.2857	-1.0000	0.2975	0.2383
80	1.0000	0.1510	0.0228	-1.0000	-0.0955	-0.0077	-1.0000	-0.0940	-0.0073	-3.0000	-0.3199	-0.0328
81	-1.0000	-0.1569	-0.0246	-3.0000	-0.4373	-0.0639	-3.0000	-0.4343	-0.0631	-3.0000	-0.4356	-0.0635
MEM	4	4	4	2	2	2	2	2	2	0	1	1
NON	1	1	1	3	3	3	3	3	3	5	4	4
82	1.0000	0.3037	0.0922	1.0000	0.3113	0.0941	3.0000	0.5461	0.1217	3.0000	0.5435	0.1211
83	-1.0000	-0.2320	-0.0538	-3.0000	-0.5869	-0.1170	-5.0000	-0.9163	-0.1712	-7.0000	-1.2404	-0.2237
84	-1.0000	-0.5104	-0.2605	-3.0000	-0.9446	-0.3566	-5.0000	-1.2693	-0.4093	-7.0000	-1.5823	-0.4583
85	-1.0000	-0.1803	-0.0325	-3.0000	-0.4796	-0.0776	-5.0000	-0.7319	-0.1094	-7.0000	-0.9771	-0.1395
86	-1.0000	-0.1445	-0.0209	-3.0000	-0.4001	-0.0536	-5.0000	-0.6412	-0.0827	-7.0000	-0.8695	-0.1088
87	1.0000	0.1985	0.0394	3.0000	0.5053	0.0865	5.0000	0.7930	0.1279	7.0000	1.0153	0.1526
88	-1.0000	-0.1133	-0.0128	-3.0000	-0.3339	-0.0372	-5.0000	-0.5358	-0.0576	-7.0000	-0.7364	-0.0777
89	-1.0000	-0.2142	-0.0459	-3.0000	-0.5883	-0.1159	-5.0000	-0.9536	-0.1826	-7.0000	-1.2928	-0.2402
90	-1.0000	-0.3032	-0.0920	-1.0000	-0.2929	-0.0889	-3.0000	-0.5408	-0.1197	-5.0000	-0.7854	-0.1496
91	-1.0000	-0.2286	-0.0523	-3.0000	-0.5735	-0.1119	-5.0000	-0.8948	-0.1635	-7.0000	-1.2113	-0.2136
92	-1.0000	-0.2668	-0.0712	-3.0000	-0.7620	-0.1938	-5.0000	-1.2139	-0.2959	-7.0000	-1.6267	-0.3813
93	1.0000	0.1264	0.0160	1.0000	0.1375	0.0183	-1.0000	-0.0557	-0.0004	-3.0000	-0.2435	-0.0180
94	1.0000	0.4713	0.2222	-1.0000	-0.1606	0.0172	-3.0000	-0.5355	-0.0531	-5.0000	-0.8760	-0.1113
95	-1.0000	-0.1526	-0.0233	-3.0000	-0.4373	-0.0638	-5.0000	-0.7082	-0.1006	-7.0000	-0.9490	-0.1296
96	1.0000	0.1592	0.0253	3.0000	0.4451	0.0662	5.0000	0.7274	0.1061	7.0000	1.0014	0.1436
MEM	5	5	5	4	4	5	3	3	3	3	3	3
NON	10	10	10	11	11	10	12	12	12	12	12	12

KNN TIME (SEC) 15

PR49 2ND HALF

	K=1			K=3			K=5			K=7		
	SUM V	SUM V/D	SUMV/D2	SUM V	SUM V/D	SUMV/D2	SUM V	SUM V/D	SUMV/D2	SUM V	SUM V/D	SUMV/D2
DIST MAT (SEC)	16											
97	-1.0000	-0.0964	-0.0093	-1.0000	-0.0965	-0.0093	-3.0000	-0.2715	-0.0246	-5.0000	-0.4444	-0.0396
98	1.0000	0.1514	0.0229	1.0000	0.1418	0.0208	3.0000	0.3520	0.0429	1.0000	0.1491	0.0223
99	1.0000	0.1096	0.0120	3.0000	0.2920	0.0287	3.0000	0.2916	0.0286	5.0000	0.4578	0.0424
100	-1.0000	-0.1313	-0.0172	1.0000	0.0831	0.0059	1.0000	0.0809	0.0054	1.0000	0.0814	0.0055
101	1.0000	0.1172	0.0137	1.0000	0.1174	0.0138	-1.0000	-0.1101	-0.0121	-3.0000	-0.3314	-0.0366
102	1.0000	0.1403	0.0197	1.0000	0.1520	0.0226	-1.0000	-0.0826	-0.0049	-1.0000	-0.0812	-0.0046
103	1.0000	0.1317	0.0173	1.0000	0.1174	0.0140	-1.0000	-0.0937	-0.0083	-1.0000	-0.0947	-0.0085
MEM	5	5	5	6	6	6	3	3	3	3	3	3
NON	2	2	2	1	1	1	4	4	4	4	4	4
104	1.0000	0.1403	0.0197	-1.0000	-0.1083	-0.0112	-3.0000	-0.3468	-0.0397	-3.0000	-0.3474	-0.0398
105	-1.0000	-0.1194	-0.0143	-3.0000	-0.3401	-0.0386	-3.0000	-0.3425	-0.0391	-5.0000	-0.5469	-0.0600
106	-1.0000	-0.1950	-0.0380	-3.0000	-0.5745	-0.1101	-3.0000	-0.5775	-0.1111	-5.0000	-0.9197	-0.1697
107	-1.0000	-0.1903	-0.0362	-1.0000	-0.1797	-0.0326	-3.0000	-0.5017	-0.0845	-5.0000	-0.8135	-0.1331
108	-1.0000	-0.1450	-0.0210	-3.0000	-0.4120	-0.0567	-5.0000	-0.6713	-0.0903	-7.0000	-0.9276	-0.1231
109	-1.0000	-0.1684	-0.0284	-3.0000	-0.5001	-0.0834	-5.0000	-0.8240	-0.1359	-7.0000	-1.1347	-0.1841
110	-1.0000	-0.1713	-0.0293	-3.0000	-0.5049	-0.0850	-5.0000	-0.8349	-0.1394	-7.0000	-1.1507	-0.1893
111	-1.0000	-0.1201	- 0144	-3.0000	-0.3386	-0.0383	-3.0000	-0.3411	-0.0388	-5.0000	-0.5489	-0.0604
112	-1.0000	-0.1292	-0.0167	-1.0000	-0.1326	-0.0174	-1.0000	-0.1294	-0.0168	-1.0000	-0.1324	-0.0174
113	-1.0000	-0.1028	-0.0106	-3.0000	-0.3041	-0.0308	-5.0000	-0.4995	-0.0499	-7.0000	-0.6919	-0.0684
114	1.0000	0.1466	0.0215	3.0000	0.4112	0.0565	5.0000	0.6556	0.0865	7.0000	0.8692	0.1093
115	1.0000	0.1146	0.0131	3.0000	0.3357	0.0376	5.0000	0.5464	0.0598	3.0000	0.3464	0.0398
116	-1.0000	-0.3294	-0.1085	-1.0000	-0.3213	-0.1059	1.0000	-0.0263	-0.0624	1.0000	-0.0115	-0.0584
117	1.0000	0.1621	0.0263	-1.0000	-0.1522	-0.0231	-1.0000	-0.1531	-0.0234	-1.0000	-0.1523	-0.0232
118	-1.0000	-0.2336	-0.0546	-3.0000	-0.6395	-0.1370	-5.0000	-0.9755	-0.1935	-7.0000	-1.2948	-0.2444
119	-1.0000	-0.1796	-0.0323	-3.0000	-0.5187	-0.0898	-5.0000	-0.8301	-0.1384	-7.0000	-1.1187	-0.1801
120	1.0000	0.1104	0.0122	1.0000	0.1081	0.0117	3.0000	0.3008	0.0303	1.0000	0.1167	0.0134
121	1.0000	0.1224	0.0150	1.0000	0.1217	0.0148	3.0000	0.3427	0.0392	1.0000	0.1238	0.0153
122	-1.0000	-0.1875	-0.0352	-3.0000	-0.5461	-0.0994	-5.0000	-0.8818	-0.1558	-7.0000	-1.1970	-0.2055
123	1.0000	0.1305	0.0170	3.0000	0.3897	0.0506	3.0000	0.3827	0.0489	3.0000	0.3693	0.0458
124	1.0000	0.1714	0.0294	3.0000	0.5111	0.0871	3.0000	0.4965	0.0824	3.0000	0.4888	0.0804
125	-1.0000	-0.2607	-0.0680	-3.0000	-0.7100	-0.1689	-5.0000	-1.1426	-0.2625	-7.0000	-1.5553	-0.3477
MEM	8	8	8	6	6	6	7	6	6	7	6	6
NON	14	14	14	16	16	16	15	16	16	15	16	16

KNN TIME (SEC) 20

```
BEGIN
   REAL IMEAS,I,N,N1,N2,C1,C2,CCPR1,CCPR2,IAB,MER,V1,V2,W1,W2,LINECOUNT ;
   REAL IMPC,P1LJ,P2LN,D1,D2,IMAX,MMAX,P1,P2,IFLG,OAPR,IAN,NANAL,ITRPR ;
   REAL WETPEN ;
   ARRAY NTOT,NMEM,MEMMISCL,NONMISCL[0:90,0:10], IDENT[0:90], AV[0:10] ;
   ARRAY HT[0:90], CMT[0:90,0:5], TITLE[0:12] ;
   FILE LP(KIND=PRINTER),CR(KIND=READER) ; LABEL LEOF ;

   PROCEDURE ICALC(N,N1,N2,C1,C2,P1,P2,IAB,MER) ;
   %        CALCULATE I,FIG OF MERIT
   REAL N,N1,N2,C1,C2,P1,P2,IAB,MER ;
   BEGIN
      DEFINE IX(A,B,C)= (IF(A)>0 THEN (A)*LOG((A)/(B)/(C)) ELSE 0 ) # ;
      REAL PM,PN,P1M,P2M,P1N,P2N ;
      P1:=N1/N ; P2:=N2/N ;
      PM:=(C1+N2-C2)/N ; PN:=(N1-C1+C2)/N ;
      P1M:=C1/N ; P2M:=(N2-C2)/N ; P1N:=(N1-C1)/N ; P2N:=C2/N ;
      IF CCPR1+CCPR2<=1 THEN IAB:=MER:=0 ELSE
      BEGIN
         IAB:=3.32193*(IX(P1M,P1,PM)+IX(P2M,P2,PM)+IX(P1N,P1,PN)+IX(P2N,P2,PN)) ;
         MER:=-0.30103*IAB/(P1*LOG(P1)+P2*LOG(P2)) ;
      END ;
   END ;

   PROCEDURE ALLCALC ;
   %        CLASSIFICATION MEASURES CALCULATIONS
   BEGIN
      C1:=N1-MEMMISCL[IAN,ITRPR] ; C2:=N2-NONMISCL[IAN,ITRPR] ;
      %        CLASS CONDITIONAL,OVERALL,ETC PRS
      IF N1>0 THEN CCPR1:=C1/N1 ELSE
      BEGIN
         CCPR1:=0 ; WRITE(LP,<X4,"CCPR1 SET TO ZERO",/>) ; LINECOUNT:=*+1 ;
      END ;
      IF N2>0 THEN CCPR2:=C2/N2 ELSE
      BEGIN
         CCPR2:=0 ; WRITE(LP,<X4,"CCPR2 SET TO ZERO",/>) ; LINECOUNT:=*+1 ;
      END ;
      IF C1+N2-C2>0 THEN P1LJ:=C1/(C1+N2-C2) ELSE
      BEGIN
         WRITE (LP, <X4,"P1LJ SET TO ZERO",/>) ; LINECOUNT:=*+1 ; P1LJ:=0 ;
      END ;
      IF C2+N1-C1>0 THEN P2LN:=C2/(C2+N1-C1) ELSE
      BEGIN
         WRITE (LP, <X4,"P2LN SET TO ZERO",/>) ; LINECOUNT:=*+1 ; P2LN:=0 ;
      END ;
      %        ORDINARY INFO GAIN I(A,B)
      ICALC(N,N1,N2,C1,C2,P1,P2,IAB,MER) ;
      %        OVERALL PR AND IMPROV OVER MOST POP CATEG CLASSIF
      OAPR:=(C1+C2)/N ; IMPC:=OAPR-MAX(P1,P2) ;
      %        I(MAX) (REDUCED TO EQUAL NOS MEM,NON MEM)
      V1:=V2:=N/2 ;
      IF N1>0 THEN W1:=N*C1/(2*N1) ;
      IF N2>0 THEN W2:=N*C2/(2*N2) ;
      ICALC(N,V1,V2,W1,W2,D1,D2,IMAX,MMAX) ;
      AV[1]:=*+CCPR1 ; AV[2]:=*+CCPR2 ; AV[3]:=*+OAPR ; AV[4]:=*+IMPC ;
      AV[5]:=*+P1LJ ; AV[6]:=*+P2LN ; AV[7]:=*+IAB ; AV[8]:=*+MER ;
      AV[9]:=*+IMAX ;
      IF CCPR1>.999 THEN CCPR1:=.999 ; IF CCPR2>.999 THEN CCPR2:=.999 ;
      IF OAPR>.999 THEN OAPR:=.999 ; IF P2LN>.999 THEN P2LN:=.999 ;
      IF P1LJ>.999 THEN P1LJ:=.999 ; IF IAB>.999 THEN IAB:=.999 ;
      IF MER>.999 THEN MER:=.999 ; IF IMAX>.999 THEN IMAX:=.999 ;
      CASE IMEAS-1 OF
      BEGIN
         HT[IAN]:=CCPR1 ;
         HT[IAN]:=CCPR2 ;
         HT[IAN]:=OAPR ;
         HT[IAN]:=IMPC ;
         HT[IAN]:=P1LJ ;
         HT[IAN]:=P2LN ;
         HT[IAN]:=IAB ;
         HT[IAN]:=MER ;
         HT[IAN]:=IMAX ;
      END ;
      WRITE (LP,<X20,
      I2,X2,A6,I4,4(X1,F4.3),X1,F5.3,5(X1,F4.3),X10,2I4>,
      IAN,IDENT[IAN],N1,P1,CCPR1,CCPR2,OAPR,IMPC,P1LJ,P2LN,IAB,MER,IMAX,
      MEMMISCL[IAN,ITRPR],NONMISCL[IAN,ITRPR]) ;
      LINECOUNT:=*+1 ;
   END ;
```

```
PROCEDURE HISTO(ISET,NANAL,HT) ;
%         DRAW HISTOGRAMS FOR ONE SET
REAL ISET,NANAL ; ARRAY HT[*] ;
BEGIN
   ARRAY X,Y[0:10] ; REAL I ;
   AORIG(200,850-250*ISET) ;
   FOR I:=1 STEP 1 UNTIL NANAL DO
   BEGIN
      X[0]:=X[1]:=I*25 ; Y[0]:=0 ; Y[1]:=MAX(HT[I],HT[I+1])*250 ;
      ALINE(X,Y,2,0,0,100,100) ;
      X[0]:=25*I-25 ; X[1]:=25*I ; Y[0]:=Y[1]:=HT[I]*250 ;
      ALINE(X,Y,2,0,0,100,100) ;
   END ;
   IF ISET=3 THEN AEND ;
END ;


PROCEDURE HISTOAXES(NANAL,ANALLBL,IMEAS,WETPEN) ;
%         DRAW AND LABEL BOX FOR GROUP OF 3 HISTOGRAMS ;
REAL NANAL,IMEAS,WETPEN ;
ARRAY ANALLBL[*] ;
BEGIN
   POINTER PA,PB ; ARRAY X,Y[0:25],C,LBL[0:6] ; REAL IX,IY,IIX,N,I ;
   LABEL L1 ;


   PROCEDURE ORDLAB(ROW) ;
   %         1ST AND 2ND ROWS OF ANAL LABELS
   REAL ROW ;
   BEGIN
      IY:=10-35*ROW ;
      FOR I:=ROW STEP 2 UNTIL NANAL DO
      BEGIN
         IX:=25*I-33 ; C[0]:=ANALLBL[I] ; ALAB(IX,IY,C,1,2,2) ;
         PA:=POINTER(ANALLBL[I]) ;
         PB:=POINTER(LBL) ;
         IIX:=25*I-13 ;
         REPLACE PB BY PA+1 FOR 5 ;
         ALAB(IIX,IY,LBL ,5,1,2) ;
      END ;
   END ;

   AINIT(300+25*NANAL) ;
   IF WETPEN^=0 THEN
   BEGIN
      FILL LBL WITH "PLEASE USE WETPEN FOR NEXT PLOT CHEM175    " ;
      ATYPE(LBL,39) ;
   END ;
   AORIG(200,100) ;
   ABOX(0,0,1,15,NANAL*25,50,1) ;
   X[0]:=0 ; X[1]:=NANAL*25 ; Y[0]:=Y[1]:=250 ;
   ALINE(X,Y,2,0,0,100,100) ;
   Y[0]:=Y[1]:=500 ; ALINE(X,Y,2,0,0,100,100) ;
   %         LH SCALE
   FOR I:=0,1,2 DO BEGIN X[I]:=-30 ; Y[I]:=250*I-  5 ; END ;
   C[0]:="0      " ; ALINEC(X,Y,3,0,0,100,100,C[0],0,0,1,2) ;
   FILL LBL WITH 5("F2.1  ") ;
   FOR IY:=45,295,545 DO ASCALE(-40,IY,0,50,.2,.2,4,1,2,LBL,2) ;
   %         TR 76 ETC LABELS
   C[0]:="T      " ; ALAB(-105,730,C,1,2,2) ;
   C[0]:="P      " ; FOR IY:=480,230 DO ALAB(-105,IY,C,1,2,2) ;
   C[0]:="R      " ; FOR IY:=730,480,230 DO ALAB(-85 ,IY,C,1,1,2) ;
   C[0]:="76     " ; ALAB(-65 ,730,C,2,1,2) ;
   C[0]:="20     " ; ALAB(-65,480 ,C,2,1,2) ;
   C[0]:="49     " ; ALAB(-65,230 ,C,2,1,2) ;
```

```
            TYPE OF CLASSIF MEASURE LABEL
      CASE IMEAS-1 OF
      BEGIN
         BEGIN C[0]:="P      "  ; FILL LBL WITH "1      " ; N:=1 ; END ;
         BEGIN C[0]:="P      "  ; FILL LBL WITH "2      " ; N:=1 ; END ;
         BEGIN C[0]:="P      "  ; FILL LBL WITH "TOT    " ; N:=3 ; END ;
         BEGIN C[0]:="I      "  ; FILL LBL WITH "MPC    " ; N:=3 ; END ;
         BEGIN C[0] :="P      " ; FILL LBL WITH "(1$2$J)      " ; N:=7 ; END ;
         BEGIN C[0]:="P      "  ; FILL LBL WITH "(2$2$N)      " ; N:=7 ; END ;
         BEGIN C[0]:="I      "  ; FILL LBL WITH "(A,B) " ; N:=5 ; END ;
         BEGIN C[0]:="F      "  ; FILL LBL WITH "IGURE OF MERIT  " ; N:=14 ; END ;
         BEGIN C[0]:="I      "  ; FILL LBL WITH "MAX    " ; N:=3 ; END ;
      END ;
      ALAB(-50,365-5*N,C    ,1,2,4) ;
      ALAB(-50,385-5*N,LBL,N,1,4) ;
      %        ORDINATE LABELS - ANALYSIS IDENTS
      Y[0]:=-5 ; Y[1]:=-35 ;
      FOR I:=0 STEP 1 UNTIL NANAL DIV 2-1 DO
      BEGIN X[0]:=X[1]:=37+50*I ; ALINE(X,Y,2,0,0,100,100) ; END ;
      ORDLAB(1) ; ORDLAB(2) ;
      %     HEADING AT TOP OF HISTOGRAM
      PA:=POINTER(TITLE[12])+5 ;
      FOR I:=1 STEP 1 UNTIL 77 DO
      IF PA EQL " " THEN PA:=*-1 ELSE
      BEGIN
         IX:=12.5*NANAL+10*I-780 ;
         IY:=770 ;
         I:=79-I ;
         ALAB(IX,IY,TITLE,I,2,2) ;
         GO TO L1 ;
      END ;
      L1: ;
   END ;

   %        ********************

   LINECOUNT:=35 ;
   READ (CR,<13A6>,TITLE) ;
   FOR IAN:=1 STEP 1 UNTIL 1000 DO
   READ (CR,<A6,12I3,6A6>,
   IDENT[IAN],FOR I:=1,2,3 DO [NTOT[IAN,I],NMEM[IAN,I],
   MEMMISCL[IAN,I],NONMISCL[IAN,I]],CMT[IAN,*]) [::LEOF] ;
   LEOF : NANAL:=IAN-1 ;
   READ (CR,<40I2>,IMEAS,WETPEN) ;
   %        BMD07M CLASSIFS ON TR,PR96,PR125 SETS

   WRITE (LP,<13A6>, TITLE) ;
   HISTOAXES(NANAL,IDENT,IMEAS,WETPEN) ;
   FOR ITRPR:=1,2,3 DO
   BEGIN
      BEGIN
         IF ITRPR=3 THEN WRITE(LP,</>)   ELSE WRITE (LP[SKIP 1]) ;
         WRITE (LP,<    X53,"IMPROV">) ;
         WRITE (LP,<X41,"P    P    P    MOST      P(2!N)      FIG",
         X15,"NO.MISCL">) ;
         WRITE (LP,<X31,"MEM P(1)   1    2   TOT  POP P(1!J)    I(A,B) MER IMAX",
         X11,"MEM NON",//>) ;
      END ;
      FILL HT WITH 27(0) ;
      FOR I:=1,2,3,4,5,6,7,8,9 DO AV[I]:=0 ;
      FOR IAN:=1 STEP 1 UNTIL NANAL DO IF NTOT[IAN,ITRPR]>0 THEN
      BEGIN
         N:=NTOT[IAN,ITRPR] ; N1:=NMEM[IAN,ITRPR] ; N2:=N-N1 ; ALLCALC ;
      END ;
      FOR I:=1,2,3,4,5,6,7,8,9 DO
      BEGIN AV[I]:=AV[1]/NANAL ; IF AV[I]>0.999 THEN AV[I]:=0.999 ; END ;
      WRITE(LP,</,X24,"AV.",X12,3(X1,F4.3),X1,F5.3,5(X1,F4.3)>,
      FOR I:=1,2,3,4,5,6,7,8,9 DO AV[I]) ;

      HISTO(ITRPR,NANAL,HT) ;
   END
END .
```

Appendix III

DETAILED RESULTS OF ANALYSES

The performance of programs NUCL and MOLION (chapter 4) on each of the 125 spectra is reported in section III.1. Section III.2 contains the weight vectors computed for each of the twenty-one structural categories by the statistical discriminant function analysis of chapter 6, and for each of the eleven convergent categories by the learning machine approach of chapter 7. The $\leqslant 82$ dimension means of chapter 8 are reported in section III.3. The mass positions used for the k-nearest neighbour approach of chapter 9 are also included. Only the most efficacious variant of each of the pattern recognition methods is recorded.

III.1 Molecular and Base Weight Determination

Results of the molecular weight determination and base weight determination procedures of chapter 4 are presented for each of the 125 spectra in table III.1. Abbreviations are as follows:

MW : molecular weight determination,

B : base weight determination,

Spect : spectrum number as in appendix I,

Molion: molecular weight determination by program MOLION,

M : molecular weight determination by losses from M procedure of program NUCL,

B+S : molecular weight determination by losses from base + sugar procedure of program NUCL,

As Nucl: base weight determination as a C-N bonded nucleoside by program NUCL,

As CG : base weight determination as a C-C bonded carbon glycoside by program NUCL,

1 : correct value ranked as the most likely candidate,

5 : correct value ranked amongst the top five candidates,

nc : no candidates postulated,

nd : correct value not detected.

Table III.1: [Overleaf] Molecular and base weight determination.

Spect	MW			B	
	Molion	M	B+S	As Nucl	As CG
1	1	1	1	1	nc
2	nd	5	1	1	nd
3	1	1	1	1	nd
4	1	1	1	1	nd
5	nd	5	5	nd	5
6	1	1	1	1	nd
7	1	nd	5	1	nd
8	1	1	1	1	nd
9	1	1	1	1	nd
10	1	5	1	1	nd
11	1	1	nd	1	nd
12	5	1	nd	1	nd
13	1	nd	nd	1	nd
14	1	5	nd	1	nd
15	1	nd	nd	1	nd
16	5	1	1	1	nd
17	1	nc	nd	1	nd
18	1	nc	nc	1	nd
19	nd	5	nc	5	nd
20	1	5	nc	nd	nd
21	1	nd	nc	nd	nd
22	1	1	nd	nd	nd
23	1	nd	nd	nd	5
24	nd	nd	1	5	5
25	1	1	1	1	nd
26	5	1	1	5	nd
27	1	1	1	1	nd
28	1	1	nc	nd	nc
29	5	1	1	1	nd
30	1	1	1	5	nd
31	5	1	1	1	nd
32	5	1	1	1	nd

Spect	M			B	
	Molion	M	B+S	As Nucl	As CG
33	1	1	1	1	nd
34	1	nc	nc	nd	nc
35	1	nc	nd	nd	nc
36	5	1	1	1	nc
37	1	1	1	1	nd
38	1	1	1	1	nd
39	1	1	1	1	nd
40	5	nd	nd	nd	nd
41	5	1	1	1	nd
42	5	1	1	1	nd
43	5	5	5	1	1
44	5	1	nd	nd	1
45	5	1	nd	nd	1
46	nd	5	5	1	1
47	1	nd	nc	nd	1
48	nd	nd	nc	nd	1
49	1	nd	nc	nd	1
50	nd	5	5	5	nd
51	1	1	1	1	nd
52	1	1	nd	1	nd
53	1	1	1	5	nd
54	nd	nd	nc	5	nd
55	1	5	1	5	5
56	1	5	nd	nd	1
57	nd	nd	5	5	nd
58	1	1	1	1	1
59	5	1	nc	nd	nd
60	1	5	nd	5	nd
61	nd	5	1	5	5
62	nd	1	1	1	1
63	5	5	5	5	nd
64	5	1	1	1	nc

Spect	MW			B	
	Molion	M	B+S	As Nucl	As CG
65	5	5	5	5	nd
66	5	5	5	5	nd
67	5	5	5	5	nd
68	5	nd	nd	nd	nd
69	1	1	1	1	nd
70	1	1	1	1	nc
71	1	1	1	1	nd
72	1	1	1	1	nd
73	1	1	1	1	nd
74	1	1	1	1	nd
75	1	1	nc	1	nd
76	1	1	nc	1	nd
77	1	1	nc	5	nd
78	1	1	nc	5	nd
79	1	nc	nc	1	nc
80	nd	nc	nc	5	nd
81	5	1	1	1	nc
82	1	1	1	5	nd
83	1	1	1	5	nd
84	5	1	nd	nd	1
85	1	nc	1	5	nc
86	1	1	nc	nd	nd
87	1	5	1	5	nd
88	1	nd	1	5	nd
89	nd	5	nd	nd	nd
90	5	1	1	1	nc
91	1	1	1	5	nd
92	5	1	1	1	nd
93	5	5	1	1	nd
94	5	nc	1	1	nd

Spect	MW			B	
	Molion	M	B+S	As Nucl	As CG
95	1	1	1	1	nd
96	5	1	1	1	nd
97	5	1	1	1	nd
98	5	1	5	1	nd
99	1	1	nd	1	nd
100	nd	nd	nd	nd	nd
101	1	1	nd	1	nd
102	1	1	nc	nd	nd
103	1	1	5	nd	nd
104	1	1	1	1	nd
105	1	1	1	1	nd
106	1	1	1	1	nd
107	5	1	1	1	1
108	1	1	5	1	nd
109	1	5	nc	nd	nc
110	1	5	1	1	nc
111	1	1	1	1	nd
112	nd	nd	nd	nd	nd
113	5	1	1	1	1
114	5	nd	5	nd	5
115	5	5	5	1	nd
116	1	5	1	5	nd
117	1	1	1	1	nd
118	1	1	1	1	nd
119	1	nd	nd	1	nd
120	nd	nd	nd	5	nd
121	1	nd	nc	5	nd
122	nd	nd	nc	nd	nd
123	nd	nc	nd	5	nd
124	5	1	5	1	nd
125	1	nc	1	5	nd

Table III.1 (Cont.)

III.2 Statistical Discriminant Function Analysis and Learning Machine Approach

Augmented 24-dimension weight vectors for each of the twenty-one structural categories, computed by the statistical linear discriminant function analysis of chapter 6, are reproduced in table III.2. The weight vectors calculated by the learning machine approach of chapter 7 using the same m/z values, for those eleven structural categories which converged, are also reproduced. Autoscaled logarithmic data used in both analyses, with no deadzone in the learning machine approach.

Table III.2: [Overleaf] Augmented 24-dimension weight vectors for statistical linear discriminant function analysis (SLDF) and learning machine approach (LMA).

	CT11		CT12		CT15		
	m/z	SLDF	m/z	SLDF	m/z	SLDF	LMA
1	108	-.326	112	- .718	112	- .793	-.064
2	111	-.490	117	- .581	117	- .671	-.025
3	112	-	133	- .717	120	.258	-.008
4	117	-.749	134	.743	121	.946	.024
5	119	-.115	136	- .940	125	- .822	-.035
6	120	.385	160	1.05	133	- .281	-.032
7	127	.216	163	2.27	160	.999	.079
8	133	-.775	164	-1.10	164	-1.26	-.006
9	141	-.201	165	- .621	166	- .709	-.026
10	163	.304	170	- .386	169	- .292	-.047
11	165	-.526	171	.126	170	- .545	.003
12	169	-.387	179	- .621	171	- .094	-.056
13	170	-.395	180	- .603	178	.900	-.013
14	171	.037	183	- .653	185	- .219	-.022
15	179	-.425	185	- .788	190	.315	-.002
16	185	.119	190	1.13	194	- .740	-.038
17	194	-.528	206	.294	218	.550	.045
18	202	.303	218	1.21	219	- .907	-.090
19	218	.295	219	.611	225	.882	.039
20	225	.068	225	1.87	228	- .255	-.003
21	228	-.625	228	- .374	248	2.32	.140
22	240	-.373	248	1.60	266	- .763	-.009
23	266	.393	280	.508	280	.150	.010
24	281	.072	316	1.72	316	1.20	.092
Const		-2.12		-4.51		-3.64	-.027

	OT5		OT6		C6		
	m/z	SLDF	m/z	SLDF	m/z	SLDF	LMA
1	111	-.113	112	- .194	117	-1.12	-.005
2	112	.281	115	- .169	119	-1.12	-.059
3	113	.332	117	-1.35	120	1.39	.125
4	115	.165	125	.805	125	-1.12	-.025
5	121	.838	126	- .775	126	-	.025
6	126	.602	127	1.71	127	-1.47	-.064
7	127	.374	133	.624	133	- .632	.009
8	135	-.490	135	- .122	134	1.26	.074
9	148	-.205	139	1.51	136	-1.08	-.045
10	151	.666	141	.326	160	1.22	.040
11	155	-.419	148	- .223	163	1.12	.090
12	164	-.331	154	- .115	164	-1.20	-.002
13	169	.370	164	- .507	169	- .538	-.047
14	171	.309	169	.158	171	- .287	-.031
15	177	-.228	171	.759	180	-1.33	-.047
16	179	-	185	.730	190	.981	.022
17	211	-.411	211	- .502	211	1.49	-.016
18	218	.198	220	1.38	218	1.42	.040
19	219	.458	221	.242	219	- .215	.024
20	220	.160	228	1.06	227	-1.33	-.014
21	225	-.496	250	-1.13	248	1.75	.038
22	228	.154	258	1.00	250	- .057	.013
23	249	-.451	267	- .116	280	.601	.030
24	316	.434	316	1.68	316	1.37	.052
Const		-.218		-4.09		-4.65	.085

Table III.2 (Cont.)

	C7			C8		
	m/z	SLDF	LMA	m/z	SLDF	LMA
1	108	- .439	-.007	112	-1.77	-.002
2	112	- .636	-.007	120	3.18	.058
3	120	1.09	.068	152	-2.71	-.005
4	121	1.03	.052	160	3.18	.081
5	126	- .614	-.062	164	-1.76	.004
6	134	- .692	.040	166	-3.22	-.027
7	135	-1.15	.001	169	.830	.012
8	139	- .953	.001	176	- .173	-.004
9	149	- .896	.001	178	-1.26	.007
10	160	1.72	.072	190	1.04	.035
11	163	2.64	.054	191	-1.11	-.038
12	166	-1.36	-.020	129	-1.12	.024
13	169	.358	-.005	194	- .126	-.014
14	171	-.417	-.007	202	2.29	-.003
15	185	-.099	.004	218	1.19	.058
16	190	1.86	.046	219	.405	.033
17	191	- .474	-.048	220	3.77	.030
18	194	- .374	-.019	225	2.70	.051
19	218	1.43	.048	228	- .460	.026
20	219	.712	.029	232	4.00	.028
21	225	1.21	.037	248	4.08	.074
22	228	- .379	.030	280	.909	.046
23	248	2.01	.082	281	1.29	.024
24	316	1.56	.153	316	5.02	.174
Const		-4.75	.021		-10.9	-.014

Table III.2 (Cont.)

	C10			O1		
	m/z	SLDF	LMA	m/z	SLDF	LMA
1	110	- .309	-.021	108	.845	.048
2	112	- .757	-.028	109	- .161	.069
3	120	2.46	.054	112	.856	.108
4	126	- .467	-.028	115	.353	.055
5	133	.199	.003	127	1.34	.065
6	139	- .671	-.010	134	- .874	-.006
7	152	-1.44	-.024	135	-1.71	-.040
8	160	3.29	.135	141	.997	.067
9	164	- .958	-.042	148	.450	.043
10	166	-1.84	-.033	151	1.93	.040
11	169	.742	-.028	152	- .589	.014
12	178	.162	-.010	155	- .460	.008
13	190	1.24	.027	163	- .202	-.020
14	191	.689	-.017	164	-1.00	-.048
15	192	- .340	-.007	169	.927	.059
16	211	1.21	.039	177	- .387	-.015
17	219	-2.04	-.106	201	.867	.031
18	220	3.00	.047	206	.267	-.005
19	225	2.52	.066	219	1.65	.058
20	232	2.59	.028	221	1.36	.026
21	248	4.79	.101	248	-	-.015
22	280	.477	-.003	251	- .083	-.005
23	281	- .329	.025	267	- .437	-.018
24	316	2.86	.034	269	- .609	-.002
25				316	1.12	.076
Const		-8.00	-.041		-3.59	.107

Table III.2 (Cont.)

	O2			N4		
	m/z	SLDF	LMA	m/z	SLDF	LMA
1	112	.106	.061	111	- .396	-.006
2	113	.282	.059	112	-2.13	-.039
3	115	- .185	.056	113	.381	-.011
4	120	1.07	.050	115	- .925	-.008
5	126	1.02	.057	125	- .953	-.009
6	127	1.71	.077	126	-1.88	-.024
7	133	1.32	.045	127	- .589	-.017
8	134	- .565	.010	133	.495	.002
9	136	.245	.041	135	.517	.094
10	139	1.37	.031	136	-	.003
11	141	.514	.045	139	.398	-.005
12	163	.116	.005	148	.271	.115
13	164	-1.26	-.068	152	.869	.007
14	165	- .364	.006	163	.139	.095
15	169	.578	.066	164	.466	-.005
16	171	.764	.057	169	-.889	-.005
17	183	.090	-.002	171	-1.12	-.013
18	185	.747	.046	184	-1.66	-.011
19	211	- .948	-.011	185	- .517	-.016
20	218	2.78	.094	193	.714	.018
21	228	2.17	.058	218	-1.98	-.029
22	232	-1.03	-.055	227	.354	-.013
23	267	- .568	-.018	228	-1.07	-.012
24	316	.938	.054	257	.391	.001
Const		-5.81	.022		-3.25	.163

Table III.2 (Cont.)

	N5		NC6		OC2		
	m/z	SLDF	m/z	SLDF	m/z	SLDF	LMA
1	112	-.365	108	.802	112	5.86	.088
2	115	-.606	113	- .248	113	-1.04	-.002
3	125	-.144	115	-1.02	125	3.54	.036
4	126	-.765	117	- .762	126	1.31	.067
5	127	.108	119	- .378	127	5.15	.074
6	135	-.223	127	- .239	133	1.37	.074
7	136	.813	133	-	136	-1.55	-.042
8	139	-.442	134	.201	140	-3.83	-.023
9	148	.485	135	1.69	146	- .463	.006
10	154	-.419	136	-1.36	148	- .242	-.038
11	164	.705	139	- .957	151	2.48	.023
12	169	-.319	141	- .478	163	- .970	-.003
13	171	-.214	146	- .606	164	- .408	.020
14	178	.279	148	.357	168	1.92	.029
15	179	-.792	155	- .769	169	2.67	.080
16	183	-.295	164	.500	171	.743	-.027
17	190	.277	166	- .852	185	.264	.060
18	191	-.091	169	- .144	193	-1.04	.017
19	202	-.419	171	- .143	207	.250	.029
20	208	.799	178	.323	211	-1.76	-.066
21	227	-.522	179	- .197	218	- .761	.014
22	228	-.262	232	- .499	219	-	.044
23	248	.560	248	.458	227	.752	.079
24	258	-.081	250	.429	228	3.27	.077
Const		-2.65		-2.92		-12.4	-.047

Table III.2 (Cont.)

	Pur		Pyr			Adn	
	m/z	SLDF	m/z	SLDF	LMA	m/z	SLDF
1	108	.078	111	-2.10	-.038	108	.226
2	109	.781	112	4.37	.068	113	-.305
3	111	- .285	113	-	.024	115	-.044
4	112	- .191	115	.632	.006	117	.044
5	115	- .148	125	2.84	.048	119	.350
6	125	- .334	126	1.49	.029	121	.159
7	126	-1.12	127	6.43	.057	125	-.374
8	127	- .069	133	1.71	.089	126	-.427
9	136	.276	135	- .854	.002	127	-.304
10	141	.330	136	- .432	-.032	133	-.371
11	148	.508	146	1.72	.063	134	.375
12	169	- .401	148	- .878	-.043	135	.131
13	170	- .559	151	2.01	.063	151	-.375
14	171	- .756	164	.175	.033	154	-.336
15	177	- .375	168	3.03	-.002	166	-
16	179	- .151	171	1.12	-.067	169	-.424
17	185	- .104	185	.393	.005	170	-.558
18	191	- .624	193	-1.52	-.069	179	-.550
19	207	- .075	207	-1.01	.024	183	.428
20	211	.358	208	-4.59	-.062	185	-.796
21	218	- .547	211	-2.59	-.040	207	-.289
22	219	- .125	218	3.24	.024	218	-1.32
23	228	- .850	219	- .214	.031	228	-.353
24	257	- .471	240	1.44	.017	232	.327
Const		-2.26		-14.1	-.112		-2.37

Table III.2 (Cont.)

	AN6			Asug		S133	
	m/z	SLDF	LMA	m/z	SLDF	m/z	SLDF
1	108	- .167	.012	112	-.551	109	2.29
2	111	- .068	.013	113	.149	110	- .931
3	112	- .398	-.027	126	-.555	117	- .919
4	120	.742	.087	127	-.459	125	.514
5	126	- .618	-.040	133	-.370	133	2.77
6	127	- .895	-.066	136	.606	141	.464
7	133	- .502	-.020	149	-.625	146	- .922
8	134	.207	.056	151	-.418	148	1.84
9	135	-1.25	-.015	155	.476	149	.490
10	160	.673	.073	162	-.347	160	.385
11	163	.383	.043	164	-	162	1.37
12	169	- .534	-.102	168	-.133	164	-1.46
13	170	- .550	-.008	169	-.595	171	- .960
14	171	- .347	-.047	171	-.303	179	.267
15	177	- .250	-.043	180	-	180	-
16	185	.267	-.010	185	-.045	185	.595
17	191	- .986	-.098	190	.072	190	- .469
18	194	- .594	-.047	200	.415	200	- .721
19	201	.976	.061	201	-.523	201	-1.31
20	211	.979	.053	202	-.369	202	4.05
21	218	-1.32	-.037	207	-	219	-1.19
22	227	- .248	.002	218	-.620	227	- .147
23	228	- .832	-.008	228	-.354	232	- .557
24	232	.545	.037	248	-.499	280	.779
Const		-3.09	-.032		-2.39		-4.75

Table III.2 (Cont.)

III.3 Distance from Mean and k-Nearest Neighbour Approaches

The best variant of the distance from the mean approach of chapter 8 was $\leqslant$ 82 m/z positions with autoscaled logarithmic data and zero deadzone. The 82 m/z positions are recorded in table III.3. These are the same as those used for the k-nearest neighbour approach of chapter 9. The means for class members and non members are reproduced, in the same m/z order as table III.3, for each of the twenty-one structural categories in table III.4. These have been calculated on the training set Tr76.

m/z 1-11	m/z 12-22	m/z 23-32	m/z 33-42	m/z 43-52	m/z 53-62	m/z 63-72	m/z 73-82
108	121	146	164	179	201	221	251
109	125	148	165	180	202	225	257
110	126	149	166	183	206	226	258
111	127	151	168	185	207	227	266
112	133	152	169	190	208	228	267
113	134	154	170	191	209	232	268
114	135	155	171	192	211	240	269
115	136	160	176	193	218	248	280
117	139	162	177	194	219	249	281
119	140	163	178	200	220	250	316
120	141						

Table III.3: m/z positions used in distance from mean and k-nearest neighbour analyses. Numbering same as in table III.4.

Table III.4: [Overleaf] Class member and non member means for the twenty-one structural categories. Autoscaled logarithmic data used with $\leqslant$ 82 m/z positions.

CT11				CT12			
1-41		42-82		1-41		42-82	
MEM	NON MEM	MEM	NON MEM	MEM	NON MEM	MEM	NON MEM
-.117	.190	-	-	-.341	.291	-	-
-.113	.184	-.163	.264	-.211	.180	-.163	.139
-.092	.148	-.175	.284	-.147	.126	-.184	.157
-.217	.352	-	-	-.217	.185	-.192	.163
-.181	.293	-.116	.187	-.247	.211	-.293	.250
-.153	.249	.158	-.256	-.118	.100	.307	-.262
-.099	.160	-	-	-	-	-	-
-.144	.233	.142	-.230	-.124	.106	-	-
-.234	.380	-	-	-.244	.208	-	-
-.086	.140	-.142	.231	-	-	-.182	.155
207	-.336	-	-	.289	-.246	-	-
.098	-.159	-	-	.175	-.150	.158	-.135
-.189	.306	-	-	-.179	.153	.131	-.112
-.176	.285	.166	-.269	-.125	.107	.330	-.282
-.077	.124	.084	-.136	-.117	.100	.168	-.144
-.205	.332	-.102	.165	-.205	.175	-.122	.104
.208	-.338	-	-	.398	-.340	-	-
-	-	-	-	-.177	.151	-.125	.106
-.091	.148	.120	-.194	-.209	.179	.229	-.195
-.093	.151	.141	-.228	-	-	.350	-.299
-.116	.188	.126	-.204	-.144	.123	.291	-.248
-.196	.318	-	-	-.192	.164	.140	-.120
-	-	.125	- 203	-.219	.187	.239	-.204
.157	-.254	-.122	.198	.309	-.264	-	-
-	-	-.218	.353	.189	-.161	-.254	.216
-	-	-.136	.221	-.134	.114	-.255	.217
-	-	-	-	-.194	.166	.138	-.118
-.139	.225	-.225	.364	-.158	.135	-.213	.182
-	-	.144	-.234	-.222	.190	.306	-.261
.119	-.193	.131	-.212	.226	-.193	.248	-.212
.079	-.129	.092	-.149	.173	-.148	.166	-.141
.180	-.292	.078	-.127	.363	-.310	-.149	.127
-	-	.096	-.155	-.179	.152	-	-
-.230	.373	-	-	-.227	.194	-	-
-	-	.128	-.207	-.118	.101	-	-
-.116	.189	-.092	.148	-	-	-.109	.093
-.131	.213	-	-	-.117	.100	-	-
-.179	.291	-	-	-.179	.153	-.177	.151
-.244	.396	.150	-.243	-.244	.209	.233	-.199
-	-	.140	-.227	-	-	.149	-.128
-	-	.119	-.193	-	-	.227	-.193

CT15				OT5			
1-41		42-82		1-41		42-82	
MEM	NON MEM	MEM	NON MEM	MEM	NON MEM	MEM	NON MEM
-.244	.142	.201	-.118	-.123	.160	-	-
-.242	.141	-.163	.095	.162	-.211	-	-
-.187	.109	-.184	.107	.143	-.187	.141	-.184
-.217	.127	-.194	.113	.167	-.217	-.146	.190
-.329	.192	-.293	.171	.253	-.329	.137	-.178
-.231	.135	.239	-.140	.198	-.257	-.097	.126
-.144	.084	-	-	.127	-.165	-	-
-	-	-.211	.123	.155	-.202	-.221	.288
-.256	.149	-	-	-	-	-	-
-	-	-.164	.096	-.180	.235	-.124	.162
.162	-.094	-	-	-	-	-	-
.252	-.147	.255	-.149	.095	-.124	-	-
-.217	.127	.196	-.114	.130	-.169	-	-
-.200	.117	.211	-.123	.256	-.334	-	-
-.257	.150	-	-	.243	-.316	.095	-.123
-.205	.120	-	-	.149	-.195	.130	-.169
-	-	-	-	-.229	.298	-	-
-	-	-	-	-.312	.407	-	-
-	-	-	-	-.211	.274	.095	-.124
-.200	.116	.424	-.247	-	-	.148	-.193
-.293	.171	.494	-.288	.225	-.293	.099	-.128
-.187	.109	.223	-.130	.161	-.210	.092	-.120
-.176	.103	.233	-.136	-	-	-.107	.140
.327	-.191	-	-	-.179	.234	-	-
.316	-.184	-.239	.140	-	-	.151	-.197
-	-	-.255	.149	.177	-.231	.196	-.255
-.180	.105	.141	-.082	-.137	.178	-.144	.188
-.150	.087	-.201	.117	-.100	.131	-	-
-.279	.163	.480	-.280	-.139	.181	.104	-.136
.331	-.193	-	-	-.099	.129	-.212	.276
.288	-.168	.239	-.140	-	-	-	-
.142	-.083	-	-	-	-	-.177	.231
-.180	.105	.183	-.107	-.315	.410	-	-
-.225	.131	-	-	-	-	.127	-.165
-.177	.103	-	-	-	-	-	-
-	-	-	-	-	-	-.129	.168
-.297	.173	-	-	.228	-.297	-.137	.179
-.179	.105	-	-	.138	-.179	-.141	.184
-.244	.143	.237	-.138	.188	-.244	-	-
-	-	.256	-.150	-.206	.268	-	-
-	-	.332	-.193	-.239	.311	.148	-.193

TABLE III.4 CONT.

OT6				C6			
1-41		42-82		1-41		42-82	
MEM	NON MEM	MEM	NON MEM	MEM	NON MEM	MEM	NON MEM
-.367	.180	-	-	-.329	.281	-	-
.247	-.121	-	-	-	-	-.163	.139
.344	-.169	-	-	-	-	-.184	.157
.137	-.067	-	-	-.217	.185	-.194	.166
.252	-.123	.446	-.219	-.212	.181	-.238	.203
.453	-.222	-	-	-.236	.202	.307	-.262
.292	-.143	-.230	.113	-.159	.136	.122	-.104
.413	-.202	-.240	.118	-.124	.106	-	-
-.261	.128	.200	-.098	-.244	.208	.158	-.135
-.236	.116	-.153	.075	-	-	-.182	.155
-	-	-	-	.299	-.255	-	-
-	-	-.176	.086	.175	-.150	.181	-.154
-	-	-.160	.078	-.217	.185	.151	-.129
.514	-.252	-	-	-.334	.285	.116	-.099
.550	-.270	.259	-.127	-.304	.259	-.129	.110
.405	-.198	.214	-.105	-.205	.175	-	-
-.245	.120	-	-	.398	-.340	-	-
-.533	.261	-	-	-.134	.114	-	-
-.291	.143	-	-	-.209	.179	.229	-.195
.286	-.140	.155	-.076	-.229	.195	.378	-.323
.412	-.202	.231	-.113	-.204	.174	.325	-.277
.403	-.197	-.162	.079	-.192	.164	.140	-.120
-	-	-	-	-.186	.159	.119	-.102
-.211	.103	.248	-.122	.309	-.264	-	-
-.140	.068	.219	-.107	.189	-.161	-.239	.204
-	-	.483	-.237	-	-	-	-
-	-	-.239	.117	-.194	.166	.138	-.118
-.166	.081	-	-	-.150	.128	-.213	.182
-	-	.259	-.127	-.279	.238	.328	-.280
-.193	.095	-.212	.104	.226	-.193	.248	-.212
-.249	.122	-.202	.099	.173	-.148	.166	-.141
-	-	-.302	.148	.363	-.310	-	-
-.440	.216	.220	-.108	-.179	.152	.199	-.170
-	-	.374	-.183	-.227	.194	-	-
-	-	-	-	-	-	-	-
-.153	.075	-.196	.096	-	-	-.109	.093
.546	-.268	-.165	.081	-.297	.254	-	-
-	-	-.227	.111	-.179	.153	-.132	.113
.498	-.244	-	-	-.244	.209	.233	-.199
-.189	.092	-	-	-	-	.149	-.128
-.197	.097	.336	-.165	-	-	.227	-.193

TABLE III.4 CONT.

C7				C8			
1-41		42-82		1-41		42-82	
MEM	NON MEM	MEM	NON MEM	MEM	NON MEM	MEM	NON MEM
-.299	.206	.143	-.099	-.214	.099	-.150	.069
-.211	.145	-.163	.112	-.236	.109	-.163	.075
-.187	.129	-.184	.126	-.184	.085	-	-
-.217	.150	-.194	.134	-.217	.100	-.194	.090
-.329	.227	-.293	.202	-.329	.152	-.293	.135
-.233	.161	.385	-.265	-.226	.104	.291	-.134
-.152	.104	-	-	-	-	-	-
-	-	-	-	-	-	-.290	.134
-.241	.166	-	-	-.222	.103	-.156	.072
-	-	-.173	.119	-	-	-.198	.092
.263	-.181	-	-	.249	-.115	-	-
.220	-.151	.208	-.143	-	-	.322	-.149
-.217	.150	.178	-.123	-.217	.100	.313	-.144
-.334	.230	.172	-.118	-.334	.154	.189	-.087
-.301	.207	-.129	.089	-.301	.139	-	-
-.205	.141	-	-	-.205	.095	-	-
.258	-.178	-	-	-.226	.104	-	-
-.166	.114	-	-	-	-	-	-
-.171	.118	.284	-.195	-	-	.379	-.175
-.226	.156	.438	-.302	-.185	.086	.676	-.312
-.293	.202	.412	-.284	-.293	.135	.633	-.292
-.189	.130	.179	-.123	-.183	.085	.296	-.137
-.181	.125	.161	-.111	-.167	.077	.270	-.125
.402	-.277	-	-	.398	-.184	-	-
.254	-.175	-.239	.165	.375	-.173	-.239	.110
-	-	-.255	.175	-.249	.115	-.255	.118
-.190	.131	.187	-.129	-.224	.104	.323	-.149
-.150	.103	-.207	.143	-.150	.069	-.192	.088
-.279	.192	.406	-.280	-.279	.129	.653	-.301
.280	-.193	.148	-.102	.423	-.195	.253	-.117
.229	-.158	.196	-.135	.366	-.169	.353	-.163
.389	-.268	-	-	-.172	.080	-	-
-.132	.091	.145	-.100	-.369	.170	.248	-.114
-.226	.156	-	-	-.182	.084	-.146	.067
-.177	.122	-	-	-.177	.082	-	-
-	-	-	-	-	-	-	-
-.297	.205	-	-	-.297	.137	-	-
-.179	.124	-.126	.087	-.179	.083	-	-
-.244	.168	.190	-.131	-.244	.113	.331	-.153
-	-	.205	-.141	-	-	.346	-.159
-	-	.281	-.193	-	-	.419	-.193

TABLE III.4 CONT.

C10				O1			
1-41		42-82		1-41		42-82	
MEM	NON MEM	MEM	NON MEM	MEM	NON MEM	MEM	NON MEM
-.183	.074	-	-	-.130	.137	-.205	.216
-.235	.096	-.163	.066	.200	-.211	.102	-.108
-.184	.075	-	-	.177	-.187	.174	-.184
-.217	.088	-.194	.079	.206	-.217	-.143	.151
-.329	.134	-.293	.119	.312	-.329	.181	-.191
-.223	.091	.344	-.140	.244	-.257	-	-
-	-	-	-	.162	-.171	-	-
-.202	.082	-.287	.117	.192	-.202	-.271	.285
-.220	.090	-.153	.062	-	-	-	-
-	-	-.192	.078	-.183	.193	-.159	.167
.304	-.124	.157	-.064	-	-	.105	-.111
-	-	.372	-.152	-.121	.128	.117	-.123
-.217	.088	.330	-.134	.166	-.175	.110	-.116
-.334	.136	.236	-.096	.317	-.334	-	-
-.299	.122	-	-	.300	- 316	.121	-.127
-.205	.084	-	-	.186	-.196	.159	-.168
-.215	.087	-	-	-.344	.363	-	-
-	-	-	-	-.350	.369	-	-
-	-	-	-	-.214	.225	.125	-.132
-.183	.074	.619	-.252	.114	-.120	.225	-.237
-.293	.119	.722	-.294	.278	-.293	.174	-.184
-.181	.074	.337	-.137	.199	-.210	.132	-.139
-.161	.066	.314	-.128	-	-	-	-
.472	-.192	-	-	-.199	.210	-	-
.438	-.179	-.239	.098	-	-	.194	-.204
-.248	.101	-.255	.104	.252	-.266	.242	-.255
-.224	.091	.258	-.105	-	-	-.135	.142
-.150	.061	-.185	.076	-	-	-	-
-.279	.114	.740	-.301	-.128	.135	.176	-.186
.480	-.195	.289	-.118	-	-	-.212	.224
.422	- 172	.405	-.165	-	-	.115	-.121
-.155	.063	-	-	-.254	.268	-.138	.145
-.353	.144	-	-	-.508	.536	.105	-.111
-.180	.073	-	-	-	-	-	-
-.177	.072	-	-	-	-	-	-
-	-	-	-	-	-	-.140	.148
-.297	.121	-	-	.282	-.297	-.103	.109
-.179	.073	-	-	.170	-.179	-.132	.139
-.244	.100	.383	-.156	.232	-.244	-	-
-	-	.402	-.164	-.203	.214	-	-
-	-	.274	-.112	-.233	.245	.184	-.193

TABLE III.4 CONT.

O2				N4			
1-41		42-82		1-41		42-82	
MEM	NON MEM	MEM	NON MEM	MEM	NON MEM	MEM	NON MEM
-.359	.166	-.139	.064	.172	-.396	.163	-.375
.266	-.123	-	-	-.167	.386	-	-
.366	-.169	-	-	-.187	.430	-.065	.149
.152	-.070	-	-	-.217	.501	-	-
.371	-.171	.477	-.220	-.318	.732	-.222	.511
.527	-.243	-	-	-.243	.561	.128	-.296
.313	-.145	-.228	.105	-.146	.336	.098	-.227
.438	-.202	-.273	.126	-.202	.466	.126	-.291
.179	-.083	.177	-.082	-.075	.172	.065	-.149
-.230	.106	-.149	.069	.135	-.311	.110	-.253
.167	-.077	-	-	.084	-.193	-	-
-	-	-.174	.080	.074	-.170	.080	-.185
-	-	-	-	-.150	.345	-	-
.614	-.284	-	-	-.311	.717	-	-
.678	-.313	.276	-.128	-.314	.723	-.128	.296
.430	-.198	.317	-.146	-.199	.458	-	-
-.344	.159	-	-	.149	-.344	-	-
-.529	.244	-	-	.259	-.597	-	-
-.420	.194	.310	-.143	.213	-.490	-.124	.285
.314	-.145	-	-	-.138	.318	-	-
.441	-.204	-.250	.115	-.293	.675	.106	-.243
.428	-.198	-.162	.075	-.164	.378	.070	-.162
-	-	-	-	-	-	-	-
-.217	.100	.269	-.124	.179	-.412	-.129	.297
-	-	.239	-.110	.134	-.309	-.224	.515
-	-	.552	-.255	-	-	-.255	.587
-	-	-	-	.067	-.154	-	-
-.162	.075	-	-	-	-	-.106	.244
-	-	-.172	.079	-	-	.066	-.152
-.193	.089	-.212	.098	.084	-.193	.092	-.212
-.248	.115	-.202	.093	.112	-.259	.088	-.202
-.358	.165	-.302	.139	.155	-.358	.131	-.302
-.541	.250	.238	-.110	.235	-.541	-.140	.322
-	-	-	-	.079	-.183	-	-
-	-	-	-	-	-	.095	-.219
-	-	-.196	.090	-.123	.284	.085	-.196
.591	-.273	-.160	.074	-.273	.630	.067	-.155
-	-	-.227	.105	-.179	.413	.098	-.227
.529	-.244	-	-	-.244	.563	.106	-.243
-.187	.086	-	-	.100	-.230	.121	-.278
-.193	.089	.316	-.146	128	-.296	-	-

TABLE III.4 CONT.

N5				OC2			
1-41		42-82		1-41		42-82	
MEM	NON MEM	MEM	NON MEM	MEM	NON MEM	MEM	NON MEM
.293	-.382	.270	-.352	-.363	.121	-.386	.129
-.157	.205	-.155	.202	.271	-.090	-	-
-.187	.243	-	-	-	-	-	-
-.217	.283	-.192	.250	.312	-.104	-	-
-.329	.429	-.205	.267	.953	-.318	.680	-.227
-.240	.313	.227	-.296	.733	-.244	-.296	.099
-.129	.168	-	-	.453	-.151	-	-
-.202	.264	.219	-.285	.463	-.154	-.281	.094
-.210	.274	-	-	.264	-.088	-.183	.061
-	-	.194	-.253	-.297	.099	-.199	.066
-	-	-	-	-.162	.054	-	-
.131	-.170	.099	-.130	-.170	.057	-.163	.054
-.217	.283	-	-	.456	- 152	-.178	.059
-.334	.435	-	-	.729	-.243	-	-
-.308	.402	-.135	.176	.716	-.239	.392	-.131
-.197	.257	-.091	.118	.597	-.199	-	-
-	-	-	-	-.344	.115	.203	-.068
.403	-.526	-.138	.180	-.559	.186	-.174	.058
.367	-.478	-.106	.139	-.505	.168	-.197	.066
-.232	.302	.094	-.122	-	-	-.317	.106
-.293	.382	.225	-.293	.402	-.134	-.285	.095
-.195	.254	.116	-.151	.446	-.149	-.162	.054
-.144	.188	-	-	.173	-.058	-	-
.317	-.413	-.110	.143	-.411	.137	.412	-.137
.194	-.252	-.254	.330	-.288	.096	.680	-.227
-	-	-.255	.332	.250	-.083	.764	-.255
-.104	.136	-	-	-.199	.066	-.239	.080
-.158	.206	-.127	.165	-	-	.350	-.117
-.238	.310	.157	-.205	-	-	-.234	.078
.122	-.159	-	-	-.193	.064	-.212	.071
.198	-.259	.143	-.187	-.259	.086	-.202	.067
.175	-.229	.232	-.302	-.358	.119	-.302	.101
.415	-.541	-.124	.162	-.541	.180	-	-
-.163	.213	-	-	-.177	.059	.187	-.062
-.177	.231	.168	-.219	-	-	-.219	.073
-.111	.145	.143	-.187	.380	-.127	-.196	.065
-.268	.349	-	-	.592	- 197	-	-
-.179	.234	-	-	.254	-.085	-.227	.076
-.244	.318	.112	-.146	.548	-.183	-.243	.081
.176	-.230	.214	-.278	-.230	.077	-.200	.067
-	-	-	-	-.296	.099	-.193	.064

TABLE III.4 CONT.

NC6				PUR			
1-41		42-82		1-41		42-82	
MEM	NON MEM	MEM	NON MEM	MEM	NON MEM	MEM	NON MEM
.292	-.424	.243	-.353	.192	-.294	.145	-.222
-	-	-.155	.225	-.161	.247	-.100	.153
-.098	.142	-.099	.143	-.187	.286	-.092	.141
-	-	-.168	.244	-.217	.333	.089	-.136
-	-	-.166	.241	-.318	.487	-.266	.408
-.241	.350	.204	-.296	-.241	.370	.136	-.209
-.133	.193	.121	-.176	-.135	.207	-	-
-.202	.294	.203	-.295	-.202	.310	-	-
-.226	.328	-	-	-	-	.095	-.146
-	-	.170	-.247	.176	-.270	.145	-.222
-	-	.088	-.128	.122	-.188	-	-
.117	-.170	.143	-.207	.111	-.170	-	-
-	-	-	-	-.145	.222	-	-
-.276	.400	-	-	-.311	.477	-	-
-.312	.453	-.134	.194	-.312	.479	-.137	.210
-.197	.287	-.167	.242	-.198	.303	-	-
.237	-.344	-.198	.287	-	-	-	-
.440	-.638	-.154	.223	.293	-.450	-	-
.247	-.358	-.110	.160	.299	-.458	-.163	.250
-.233	.338	-	-	-.123	.188	-	-
-.121	.175	.181	-.263	-.293	.449	.166	-.254
-.196	.284	.104	-.151	-.162	.249	.098	-.150
-.147	.213	-	-	-	-	-	-
.284	-.412	-.104	.151	.244	-.374	-.111	.171
.171	-.248	-.138	.200	.113	-.174	-.217	.332
-	-	-.131	.190	-	-	-.255	.391
-.185	.269	-	-	.119	-.183	-	-
-.105	.153	-	-	-	-	-.131	.201
-.240	.348	.131	-.190	-	-	.122	-.188
.133	-.193	.146	-.212	.101	-.155	-	-
.178	-.259	.139	-.202	.157	-.241	.105	-.161
.223	-.324	.208	-.302	.133	-.205	.143	-.219
.373	-.541	-	-	.156	-.239	-.206	.316
-.153	.222	-	-	-	-	-	-
-.177	.257	.143	-.208	-	-	.091	-.140
-	-	.114	-.166	-.115	.177	.220	-.337
-.269	.391	-	-	-.270	.414	-	-
-	-	-.090	.131	-.179	.275	-	-
-.244	.355	.168	-.243	-.244	.375	.089	-.136
.158	-.230	.192	-.278	-	-	.144	-.221
.204	-.296	-.174	.253	-.121	.185	-	-

TABLE III.4 CONT.

PYR				ADN			
1-41		42-82		1-41		42-82	
MEM	NON MEM	MEM	NON MEM	MEM	NON MEM	MEM	NON MEM
-.490	.131	-.391	.104	.181	-.263	.112	-.162
.218	-.058	-	-	-.160	.232	-.118	.171
-	-	-.184	.049	-.187	.271	-.099	.143
.411	-.110	-	-	-.217	.315	.094	-.137
1.161	-.310	.825	-.220	-.329	.478	-.293	.425
.902	-.241	-.296	.079	-.241	.350	.126	-.183
.438	-.117	-.207	.055	-.133	.193	.098	-.142
.587	-.157	-.292	.078	-.202	.294	-	-
.362	-.097	-.196	.052	-	-	-.091	.131
-.359	.096	-.253	.067	.179	-.260	-	-
-.311	.083	-	-	.158	-.229	-	-
-.170	.045	-.165	.044	.117	-.170	-	-
.433	-.115	-.188	.050	-.182	.264	-	-
.852	-.227	.198	-.053	-.334	.485	-	-
.876	-.234	.496	-.132	-.314	.456	-.134	.194
.748	-.199	-.182	.048	-.197	.287	-.124	.180
-.344	.092	.280	-.075	.237	-.344	-	-
-.580	.155	-.172	.046	.362	-.525	-	-
-.505	.135	-.197	.053	.302	-.439	-.186	.270
.174	-.046	-.317	.084	-.129	.187	-	-
.477	-.127	-.285	.076	-.293	.425	.131	-.190
-	-	-.162	.043	-.196	.284	.104	-.151
.250	-.067	-	-	-	-	-	-
-.410	.109	.387	-.103	.238	-.346	-.116	.169
-.305	.081	.553	-.147	-	-	-.216	.313
-	-	.738	-.197	-.137	.200	-.255	.370
-.276	.074	-.239	.064	.090	-.130	-	-
-	-	.464	-.124	-	-	-.169	.245
-	-	-.234	.062	-	-	.107	-.156
-.193	.051	-.212	.057	.133	-.193	.146	-.212
-.259	.069	-.202	.054	.139	-.202	.113	-.164
-.358	.095	-.302	.080	.204	-.297	.153	-.222
-.541	.144	-	-	.125	-.182	-.206	.299
-.199	.053	.231	-.062	-.094	.137	-	-
-	-	-.219	.058	-	-	.091	-.131
.484	-.129	-.196	.052	-.114	.165	-	-
.171	-.046	-.259	.069	-.269	.391	-	-
.336	-.090	-.227	.061	-.179	.260	-	-
.544	-.145	-.243	.065	-.244	.355	.168	-.243
-.230	.061	-.278	.074	-.118	.171	.153	-.223
-.296	.079	-.193	.052	-	-	-	-

TABLE III.4 CONT.

AN6				ASUG			
1-41		42-82		1-41		42-82	
MEM	NON MEM	MEM	NON MEM	MEM	NON MEM	MEM	NON MEM
-.243	.187	-	-	-	-	-	-
-.211	.162	-	-	-.211	.170	-	-
-.187	.143	-.184	.141	-.187	.151	-.184	.149
-.217	.167	.190	-.146	-.217	.176	.179	-.145
-.318	.244	-.266	.204	-.329	.266	-.293	.237
-.235	.180	.238	-.183	-.257	.208	-	-
-.156	.119	-	-	-.167	.135	-	-
-.202	.155	-	-	-.202	.164	-	-
-	-	-	-	-	-	-	-
.209	-.161	-.234	.179	-	-	.140	-.113
.338	-.259	-	-	-	-	-.213	.173
.196	-.151	.205	-.157	.129	-.105	-.230	.186
-.130	.099	.116	-.089	-.171	.138	-.227	.184
-.311	.239	.142	-.109	-.334	.271	-.211	.170
-.310	.238	-.131	.101	-.310	.251	-.155	.126
-.205	.157	-	-	-.205	.166	-.111	.089
.221	-.170	-	-	.232	-.188	-	-
-.177	.135	-	-	.277	-.224	-	-
-.165	.127	-.185	.142	.309	-.250	-.195	.158
-	-	.128	-.098	-	-	-	-
-.293	.225	.260	-.200	-.293	.237	.206	-.167
-.156	.120	.175	-.134	-.210	.170	-.121	.098
-	-	.140	-.107	-	-	-	-
.353	-.271	-	-	-.156	.126	-	-
.225	-.173	-.197	.151	-.260	.210	-.199	.161
-.231	.177	-.255	.196	-.232	.188	-.255	.206
.178	-.137	-	-	.165	-.134	-	-
.131	-.100	-	-	.122	-.099	-.140	.113
.181	-.139	.269	-.207	.168	-.136	-	-
.251	-.193	.126	-.097	-	-	-	-
.251	-.193	.182	-.140	-.170	.137	-.150	.122
.356	-.273	-	-	.139	-.112	.142	-.115
-.210	.161	-.206	.158	.169	-.137	-.206	.167
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-.175	.142	-.139	.113
-.297	.228	.134	-.103	-.297	.241	-	-
-.179	.138	-	-	-.179	.145	.135	-.110
-.244	.188	.164	-.126	-.244	.198	-	-
-	-	.175	-.135	-.161	.131	-.191	.155
-.134	.103	.118	-.091	-	-	-	-

TABLE III.4 CONT.

S133			
1-41		42-82	
MEM	NON MEM	MEM	NON MEM
.239	-.111	-	-
.456	-.211	.260	-.120
.347	-.160	.382	-.176
.152	-.070	-	-
-	-	.192	-.089
-	-	.300	-.139
.252	-.116	.283	-.130
.324	-.150	-	-
-.261	.120	-	-
.195	-.090	-	-
.177	-.082	.462	-.213
-	-	.486	-.224
.350	-.162	.465	-.215
.349	-.161	.192	-.089
-	-	-	-
.444	-.205	.172	-.079
-	-	-.146	.068
-	-	-	-
-	-	-	-
-	-	-.172	.079
.223	-.103	-.158	.073
-	-	.293	-.135
-.279	.129	-	-
.529	-.244	-	-
.606	-.280	-.268	.124
-	-	.151	-.070
-	-	.177	-.082
-	-	.351	-.162
-.153	.071	-	-
-	-	-	-
.424	-.196	.346	-.160
-	-	-	-
-	-	-	-
.261	-.120	-	-
-	-	-	-
-	-	.330	-.152
-	-	.214	-.099
-	-	-	-
.227	-.105	.317	-.146
.358	-.165	.407	-.188
.243	-.112	-.193	.089

TABLE III.4 CONT.

Appendix IV

RANDOM ASSIGNMENT CLASSIFICATION

With the burgeoning application of pattern recognition it is important to guard against improper use of the technique. For example, if the spectra studied in this work were randomly labelled as either class members or non members and the methods of chapters 6-9 applied, it would be hoped that very poor "prediction" would result as compared with that achieved in those chapters. If this were not the case, the results of chapters 6-9 would be somewhat suspect, to put it mildly. This appendix demonstrates that such a random assignment does indeed produce essentially meaningless predictions.

A 'trivial name' assignment was adopted whereby the spectra of compounds were assigned as class members if the compounds were known by a common or trivial name, and as class non members if they had solely a systematic name. This method of assignment was suggested by a similar study of Clerc et al. [321] who purported to establish a

		MEM	P(1)	P 1	P 2	P TOT	IMPROV MOST POP	P(2|N) P(1|J)		I(A,B)	FIG MER	IMAX
1	SLD	31	.408	.968	.911	.934	0.342	.882	.976	.652	.668	.679
2	DM	31	.408	.806	.933	.882	0.289	.893	.875	.451	.462	.457
3	KNN	31	.408	.999	.978	.987	0.395	.969	.999	.891	.913	.923
1	SLD	8	.400	.500	.750	.650	0.050	.571	.692	.047	.049	.049
2	DM	8	.400	.750	.917	.850	0.250	.857	.846	.361	.372	.367
3	KNN	8	.400	.999	.250	.550	-.050	.471	.999	.123	.127	.138
1	SLD	15	.306	.600	.735	.694	0.000	.500	.806	.073	.082	.084
2	DM	15	.306	.533	.676	.633	-.061	.421	.767	.028	.032	.033
3	KNN	15	.306	.800	.382	.510	-.184	.364	.813	.024	.028	.029

Table IV.1: Three pattern recognition methods applied to randomly assigned spectra. Statistical linear discriminent function analysis (SLD), distance from the mean (DM), and k-nearest neighbour (KNN) methods applied to (a) TR76, (b) Pr20, and (c) Pr49.

relationship between mass spectrum and the number of letters in a compounds' name. This deliberate absurdity was in criticism of an earlier work of Ting et al. [322] which had 'established' a relationship between mass spectrum and pharmacological activity of a set of 65 drugs, using pattern recognition methods. Clerc et al. demonstrated that their data and methods could be used to support practically any proposition whatsoever.

Table IV.1 shows the results of the statistical linear discriminant function analysis, the distance from the mean method, and the k-nearest neighbour approach on the spectra of appendix I, assigned according to type of name. The most efficacious variant of each method, as summarised in chapter 10, was used. The learning machine of chapter 7 failed to

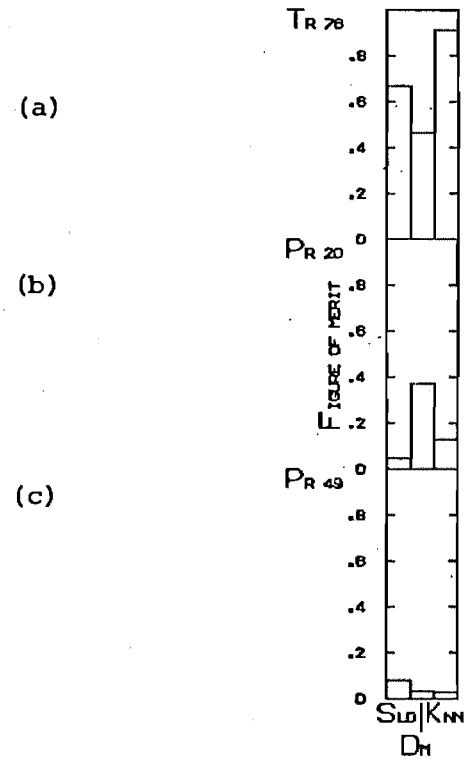


Figure IV.1: Random assignment classifications. Histograms of the results of table IV.1.

converge on the training set. As can be seen from table IV.1(c) prediction on Pr49 gave in each case a figure of merit well below 0.1. This has been graphed in figure IV.1 the bottom histogram of which shows the very poor prediction achieved.

The best distance from the mean variant gave (table 10.1) an average figure of merit over the twenty-one categories of 0.273, while the best k-nearest neighbour variant, the worst of the four methods, gave 0.125. The former at least is significantly better than any of the three results of table IV.1(c).

REFERENCES

1. J.L. Wiebers & J.A. Shapiro, Biochemistry, 16, 1044 (1977)
2. J.A. Shapiro & J.L. Wiebers, Fed. Proc. Fed. Am. Soc. Exp. Biol., 34, 607 (1975)
3. D.R. Burgard, S.P. Perone & J.L. Wiebers, Biochemistry, 16, 1051 (1977)
4. F. Caesar, Fortschr. Chem. Forsch., 39, 139 (1973)
5. A. Fozard, J.J. Franses & A.J. Wyatt in "The Applications of Computer Techniques in Chemical Research", ed. P. Hepple, p41 (Institute of Petroleum, London, 1972)
6. J.M. Gill & J. Hettinger, Comput. Chem. Biochem. Res., 2, 209 (1974)
7. J. Chromatogr. Sci., 14 (4), 165-212 (April 1976), and 14(6), 261-308 (June 1976), manuscripts from the April 1976 ACS Symposium "Impact of Microelectronics on Chromatography Automation"
8. L.R. Snyder in "Modern Practice of Liquid Chromatography", ed. J.J. Kirkland, p125 (J. Wiley & Sons, New York, 1971)
9. R.P.W. Scott, J. Chromatogr. Sci., 9, 385 (1971)
10. M. Del R. Huang & I.S. Fagerson, ibid., 13, 347 (1975)
11. W. Dulson, Anal. Chem., 49, 1279 (1977)
12. E.M. Sibley, C. Eon & B.L. Karger, J. Chromatogr. Sci., 11, 309 (1973)
13. (a) M.J. Molera, J.A. Garcia Dominguez & J. Fernandez-Biarge, ibid., 11, 538 (1973), (b) M.J. Molera, J.A. Garcia Dominguez & J. Fernandez-Biarge, ibid., 14, 299 (1976)
14. W.O. McReynolds, ibid., 8, 685 (1970)
15. (a) J.R. Mann & S.T. Preston Jr, ibid., 11, 216 (1973), (b) R.S. Henly, ibid., 11, 221 (1973)
16. (a) E.Kovats, Helv. Chim. Acta, 41, 1915 (1958), (b) L.S. Ettre, Anal. Chem., 36(8), 31A(1964), (c) G. Schomburg & G. Dielmann, J. Chromatogr. Sci., 11, 151 (1973)
17. A. Hartkopf, ibid., 12, 113 (1974)
18. L. Rohrschneider, J. Chromatogr., 22, 6 (1966)

19. J.J. Leary, J.B. Justice, S. Tsuge, S.R. Lowry & T.L. Isenhour, J. Chromatogr. Sci., 11, 201 (1973)

20. D.L. Massart, P. Lenders & M. Lauwereys, ibid., 12, 617 (1974)

21. (a) P.H. Weiner & D.G. Howery, Can. J. Chem., 50, 448 (1972), (b) D.G. Howery & P.H. Weiner, J. Chromatogr. Sci., 12, 366 (1974)

22. (a) S. Wold & K. Andersson, J. Chromatogr., 80, 43 (1973), (b) D.H. McCloskey & S.J. Hawkes, J. Chromatogr. Sci., 13, 1 (1975), (c) S. Wold, ibid., 13,525 (1975)

23. (a) T.H. Risby, P.C. Jurs & B.L. Reinbold, J. Chromatogr., 99, 173 (1974), (b) C.E. Figgins, T.H. Risby & P.C. Jurs, J. Chromatogr. Sci., 14, 453 (1976)

24. (a) D.D. Glover, Comput. Chem. Biochem. Res., 2, 101 (1974), (b) G. Michel, H.P. Sauter & A. Staubli in reference 5, p101

25. (a) R.R. Ernst in reference 5, p61, (b) D. Shaw in reference 5, p76

26. I.W. Cooley & I.W. Tuckey, Math. Comput., 19, 297 (1965)

27. D.W. Jones & T.F. Child, Adv. Magn. Reson., 8, 123 (1976)

28. A.H. Brunetti, J. Magn. Reson., 28, 289 (1977)

29. S.R. Heller, G.W.A. Milne & R.J. Feldmann, Science, 195, 253 (1977)

30. (a) D.L. Dalrymple, C.L. Wilkins, G.W.A. Milne & S.R. Heller, Org. Magn. Reson., 11, 535 (1978), (b) J. Zupan, S.R. Heller, G.W.A. Milne & J.A. Miller, Anal. Chim. Acta, 103, 141 (1978)

31. W. Voelter, G. Haas & E. Breitmaier, Chem. Ztg., 97, 567 (1973)

32. B.A. Jezl & D.L. Dalrymple, Anal. Chem., 47, 203 (1975)

33. A. Buchs, A.M. Duffield, G. Schroll, C. Djerassi, A. B. Delfino, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum & J. Lederberg, J. Am. Chem. Soc., 92, 6831 (1970)

34. S. Sasaki, Sekiyu Gakkai Shi, 14, 414 (1971) (CA 76, 3016q)

35. P.C. Jurs & T.L. Isenhour, "Chemical Applications of Pattern Recognition", (J. Wiley & Sons, New York, 1975)

36. pp124-8 of reference 35

37. (a) P.H. Weiner, E.R. Malinowski & A.R. Levinstone, J. Phys. Chem., 74, 4537 (1970), (b) P.H. Weiner & E.R. Malinowski, ibid., 75, 1207 & 3160 (1971)

38. B.R. Kowalski & C.A. Reilly, ibid., 75, 1402 (1971)

39. (a) C.L. Wilkins, R.C. Williams, T.R. Brunner & P.J. McCrombie, J. Am. Chem. Soc., 96, 4182 (1974), (b) T.R. Brunner, R.C. Williams, C.L. Wilkins & P.J. McCrombie, Anal. Chem., 46, 1798 (1974) & 47, 662 (1975), (c) C.L. Wilkins & T.L. Isenhour, ibid., 47, 1849 (1975), (d) T.R. Brunner, C.L. Wilkins, T.F. Lam, L.J. Soltzberg & S.L. Kaberline, ibid., 48, 1146 (1976), (e) C.L. Wilkins & T.R. Brunner, ibid., 49, 2136 (1977)

40. H.B. Woodruff, C.R. Snelling Jr, C.A. Shelley & M.E. Munk, ibid., 49, 2075 (1977)

41. L.P. Lindeman & J.Q. Adams, ibid., 43, 1245 (1971)

42. A.L. Burlingame, R.V. McPherson & D.M. Wilson, Proc. Nat. Acad. Sci. U.S.A., 70, 3419 (1973)

43. H.L. Suprenant & C.N. Reilley, Anal. Chem., 49, 1134 (1977)

44. J.E. Sarneski, H.L. Suprenant, F.K. Molen & C.N. Reilley, ibid., 47, 2116 (1975)

45. (a) S. Sasaki, S. Ochiai, Y. Hirota & Y. Kudo, Bunseki Kagaku, 21, 916 (1972) (CA 77, 100307 p), (b) S. Ochiai, Y. Hirota, Y. Kudo & S. Sasaki, ibid., 22, 399 (1973) (CA 79, 104438w)

46. H. Eggert & C. Djerassi, J. Am. Chem. Soc., 95, 3710 (1973)

47. R.E. Carhart & C. Djerassi, J. Chem. Soc. Perkin Trans. II, 1973, 1753

48. M. Sjostrom & U. Edlund, J. Magn. Reson., 25, 285 (1977)

49. P.C. Lauterbur, Nature (London), 242, 190 (1973)

50. (a) W.S. Hinshaw, Phys. Lett. A, 48, 87 (1974), (b) W.S. Hinshaw, J. Appl. Phys., 47, 3709 (1976), (c) E.R. Andrew, P.A. Bottomley, W.S. Hinshaw, G.N. Holland, W.S. Moore & C. Simaroj, Phys. Med. Biol., 22, 971 (1977), (d) W.S. Hinshaw, P.A. Bottomley & G.N. Holland, Nature (London), 270, 722 (1977)

51. (a) A.N. Garroway, P.K. Grannell & P. Mansfield, J. Phys. C, 7, L457 (1974), (b) P.K. Grannell & P. Mansfield, Phys. Med. Biol., 20, 477 (1975), (c) P. Mansfield & P.K. Grannell, Phys. Rev. B, 12 , 3618 (1975), (d) P. Mansfield, A.A. Maudsley & T. Baines, J. Phys. E, 9, 271 (1976), (e) P. Mansfield & A.A. Maudsley, Phys. Med. Biol., 21, 847 (1976)

52. K.E. Ekstrand, R.L. Dixon, M. Raben & C.R. Ferree, ibid., 22, 925 (1977)

53. C.W. Haigh, Annu. Rep. NMR Spectrosc., 4, 311 (1971)

54. S. Castellano & A.A. Bothner-By, J. Chem. Phys., 41, 3863 (1964)

55. Quantum Chemistry Program Exchange, Department of Chemistry, Indiana University, Bloomington, Indiana 47405, USA.

56. R. Laatikainen, J. Magn. Reson., 27, 169 (1977)

57. (a) J. Briggs, F.A. Hart & G.P. Moss, Chem. Commun., 1970, 1506, (b) M.R. Wilcott, R.E. Lenkinski & R.E. Davis, J. Am. Chem. Soc., 94, 1742 (1972)

58. I.O. Sutherland, Annu. Rep. NMR Spectrosc., 4, 71 (1971)

59. H.S. Gutowsky & C.H. Holm, J. Chem. Phys., 25, 1228 (1956)

60. B. Berglund & J. Tegenfeldt, J. Magn. Reson., 27, 315 (1977)

61. P. Moore, J. Chem. Soc. Faraday Trans. I, 72, 826 (1976)

62. (a) J.S. Mattson & A.C. McBride III, Anal. Chem., 44, 1906 (1972), (b) J.S. Mattson, ibid., 49, 470 (1977)

63. J.A. de Haseth, W.S. Woodward & T.L. Isenhour, ibid., 48, 1513 (1976)

64. W.B. Telfair, A.C. Gilby, R.J. Syrjala & P.A. Wilks Jr, Am. Lab., 8, 91 (1976)

65. T. Hirschfeld, Anal. Chem., 48, 721 (1976)

66. B.W. Tattershall, ibid., 49, 772 (1977)

67. S.T. Dunn, C.T. Foskett, R. Curbelo & P.R. Griffiths, Comput. Chem. Biochem. Res., 1, 108 (1972)

68. R. Cournoyer, J.C. Shearer & D.H. Anderson, Anal. Chem., 49, 2275 (1977)

69. R.P. Young in reference 5, p206

70. J.A. de Haseth & T.L. Isenhour, Anal. Chem., 49, 1977 (1977)

71. (a) Data base available from American Society for Testing and Materials, 1916 Rose St., Philadelphia, Pa., 19103, USA, (b) "Codes and Instructions for Wyandotte-ASTM Punched Cards", (ASTM, Philadelphia, Pa., 1964), (c) P.R. Griffiths, Anal. Chem., 46, 1206A (1974)

72. L.H. Gevantman, ibid., 44(7), 30A (1972)

73. E.C. Penski, D.A. Padowski & J.B. Bouck, ibid., 46, 955 (1974)

74. R.C. Fox, ibid., 48, 717 (1976)

75. K. Tanabe & S. Saeki, ibid., 47, 118 (1975)

76. H.B. Woodruff, S.R. Lowry, G.L. Ritter & T.L. Isenhour, ibid., 47, 2027 (1975)

77. (a) J.S. Mattson, C.S. Mattson, M.J. Spenser & F.W. Spenser, ibid., 49, 500 (1977), (b) J.S. Mattson, C.S. Matson, M.J. Spenser & S.A. Starks, ibid., 49, 297 (1977), (c) F.K. Kawahara & Y.Y. Yang, ibid., 48, 651 (1976) and discussion J.S. Mattson, ibid., 48, 2022 (1976) and F.K. Kawahara & J.F. Santner, ibid., 48, 2023 (1976)

78. (a) S.R. Lowry, H.B. Woodruff, G.L. Ritter & T.L. Isenhour, ibid., 47, 1126 (1975), (b) G.L. Ritter, S.R. Lowry, H.B. Woodruff & T.L. Isenhour, ibid., 48, 1027 (1976)

79. H.B. Woodruff, G.L. Ritter, S.R. Lowry & T.L. Isenhour, Appl. Spectrosc., 30, 213 (1976)

80. N.A.B. Gray, Anal. Chem., 47, 2426 (1975)

81. H.B. Woodruff & M.E. Munk, J. Org. Chem., 42, 1761 (1977)

82. R. Jenkins in "MTP International Review of Science. Physical Chemistry, Series One. Vol. 13. Analytical Chemistry - Part 2", eds. A.D. Buckingham & T.S. West, p127 (Butterworths, London & University Park Press, Baltimore, 1973)

83. L.S. Birks & J.V. Gilfich, Anal. Chem., 46, 360R (1974)

84. P.J. Statham, ibid., 49, 2149 (1977)

85. S.D. Rasberry & K.F.J. Heinrich, ibid., 46, 81 (1974)

86. J.W. Criss & L.S. Birks, ibid., 40, 1080 (1968)

87. M.T. Haukka & I.L. Thomas, ibid., 50, 592 (1978)

88. E. Gillam & H.T. Heal, Br. J. Appl. Phys., 3, 353 (1952)

89. K.F.J. Heinrich & S.D. Rasberry, Adv. X-Ray Anal., 17, 309 (1974)

90. M.F. Ciccarelli, Anal. Chem., 49, 345 (1977)

91. C. Palme & E. Jagoutz, ibid., 49, 717 (1977)

92. A.R. Hawthorne & R.P. Gardner, ibid., 48, 2130 (1976)

93. J. Sherman, Spectrochim. Acta, 7, 283 (1955)

94. J.W. Criss, L.S. Birks & J.V. Gilfich, Anal. Chem., 50, 33 (1978)

95. F.W. Reuter III, ibid., 47, 1763 (1975)

96. K.K. Nielson, ibid., 49, 641 (1977)

97. F.A. Mellon, Mass Spectrom., 3, 117 (1975)

98. T. Clerc & F. Erni, Fortschr. Chem. Forsch., 39, 91 (1973)

99. A.B. Delfino & A. Buchs, ibid., 39, 109 (1973)

100. M.J.E. Hewlins, Chem. Br., 12, 341 (1976)

101. G.H. Loew, R.F. Kirchner & J.G. Lawless, Org. Mass Spectrom., 11, 1158 (1976)

102. H.M. Rosenstock, Adv. Mass Spectrom., 4, 523 (1968)

103. (a) W. J. Yeager, personal communication, 1977, (b) R.G. Dromey, personal communication, 1977

104. F.W. McLafferty, R. Knutti, R. Venkataraghavan, P.J. Arpino & B.G. Dawkins, Anal. Chem., 47, 1503 (1975)

105. B. Hedfjall & R. Ryhage, ibid., 47, 666 (1975)

106. R.G. Dromey, M.J. Stefik, T.C. Rindfleisch & A.M. Duffield, ibid., 48, 1368 (1976)

107. L. Baczynskyj, D.J. Duchamp, J.F. Zieserl, M.D. Kenney & J.B. Aldrich, ibid., 48, 1358 (1976)

108. J.E. Evans & N.B. Jurinski, ibid., 47, 961 (1975)

109. I.K. Mun, R. Venkataraghavan & F.W. McLafferty, ibid., 49, 1723 (1977)

110. L.R. Crawford & J.D. Morrison, ibid., 43, 1790 (1971)

111. Yu. N. Sukharev & Yu. S. Nekrasov, Org. Mass Spectrom., 11, 1232 & 1239 (1976)

112. J. McK. Halket & R.I. Reed, ibid., 11, 881 (1976)

113 A.M. Ferguson, S.A. Gwyn, L.K. Pannell & G.J. Wright, Anal. Chem., 49, 174 (1977)

114. R. Buchi, J.T. Clerc, C. Jost, H. Koenitzer & D. Wegmann, Anal. Chim. Acta, 103, 21 (1978)

115. E. Stenhagen, S. Abrahamsson & F.W. McLafferty, "Registry of Mass Spectral Data", (J. Wiley & Sons, New York, 1974)

116. Spectra of 30, 476 different compounds available on magnetic tape from J. Wiley & Sons, 605 Third Avenue, New York, NY, 10016, USA

117. S.R. Heller, H.M. Fales & G.W.A. Milne, J. Chem. Doc., 13, 130 (1973)

118. S.R. Heller, H.M. Fales & G.W.A Milne, Org. Mass Spectrom., 7, 107 (1973)

119. R.G. Ridley in "Biochemical Applications of Mass Spectrometry", ed. G.R. Waller, p177 (J. Wiley & Sons, New York, 1972)

120. W.M. Scott & R.G. Ridley in reference 5, p148

121. D.D. Speck, R. Venkataraghavan & F.W. McLafferty, Org. Mass Spectrom., 13, 209 (1978)

122. D.H. Smith, M. Achenbach, W.J. Yeager, P.J. Anderson, W.L. Fitch & T.C. Rindfleisch, Anal. Chem., 49, 1623 (1977)

123. M. Bachiri & G. Mouvier, Org. Mass Spectrom., 11, 634 (1976)

124. B.E. Blaisdell, Anal. Chem., 49, 180 (1977)

125. (a) S.L. Grotch, ibid., 42, 1214 (1970), (b) S.L. Grotch, ibid., 43, 1362 (1971)

126. L.E. Wangen, W.S. Woodward & T.L. Isenhour, ibid., 43, 1605 (1971)

127. H.S. Hertz, R.A. Hites & K. Biemann, ibid., 43, 681 (1971)

128. S.L. Grotch, ibid., 47, 1285 (1975)

129. G. van Marlen & A. Dijkstra, ibid., 48, 595 (1976)

130. D.H. Smith, ibid., 44, 536 (1972)

131. D.H. Smith & G. Eglinton, Nature (London), 235, 325 (1972)

132. T.O. Gronneberg, N.A.B. Gray & G. Eglinton, Anal. Chem., 47, 415 (1975)

133. N.A.B. Gray & T.O. Gronneberg, ibid., 47, 419 (1975)

134. R.G. Dromey, ibid., 48, 1464 (1976)

135. S.L. Grotch, ibid., 46, 526 (1974)

136. N.A.B. Gray, ibid., 48, 1420 (1976)

137. B.A. Knock, I.C. Smith, D.E. Wright & R.G. Ridley, ibid., 42, 1516 (1970)

138. S. Farbman, R.I. Reed, D.H. Robertson & M.E.F. Silva, Int. J. Mass Spectrom. Ion Phys., 12, 123 (1973)

139. (a) F.P. Abramson, Anal. Chem., 47, 45 (1975), (b) S.C. Gates, M.J. Smisko, C.L. Ashendel, N.D. Young, J.F. Holland & C.C. Sweeley, ibid., 50, 433 (1978)

140. T.L. Isenhour, B.R. Kowalski & P.C. Jurs, Crit. Rev. Anal. Chem., 4, 1 (1974)

141. L.J. Soltzberg, C.L. Wilkins, S.L. Kaberline, T.F. Lam & T.R. Brunner, J. Am. Chem. Soc., 98, 7139 (1976)

142. T.F. Lam, C.L. Wilkins, T.R. Brunner, L.J. Soltzberg & S.L. Kaberline, Anal. Chem., 48, 1768 (1976)

143. J.B. Justice Jr & T.L. Isenhour, ibid., 46, 223 (1974)

144. J.B. Justice Jr, D.N. Anderson, T.L. Isenhour & J.C. Marshall, ibid., 44, 2087 (1972)

145. H. Rotter & K. Varmuza, Anal. Chim. Acta, 95, 25 (1977)

146. K. Varmuza, H. Rotter & P. Krenmayer, Chromatographia, 7, 522 (1974)

147. H. Rotter & K. Varmuza, Anal. Chim. Acta, 103, 61 (1978)

148. T.L. Isenhour & P.C. Jurs, Anal. Chem., 43(10), 20A (1971)

149. H. Abe & P.C. Jurs, ibid., 47, 1829 (1975)

150. B.R. Kowalski & C.F. Bender, ibid., 44, 1405 (1972)

151. W.L. Felty & P.C. Jurs, ibid., 45, 885 (1973)

152. G.S. Zander, A.J. Stuper & P.C. Jurs, ibid., 47, 1085 (1975)

153. R.J. Matthews, Aust. J. Chem., 26, 1955 (1973)

154. T.J. Stonham, I. Aleksander, M. Camp, W.T. Pike & M.A. Shaw, Anal. Chem., 47, 1817 (1975)

155. L.J. Soltzberg, C.L. Wilkins, S.L. Kaberline, T.F. Lam & T.R. Brunner, J. Am. Chem. Soc., 98, 7144 (1976)

156. J.E. Davis, A. Shepard, N. Stanford, L.B. Rogers, Anal. Chem., 46, 821 (1974)

157. R.W. Rozett & E.M. Peterssen, ibid., 47, 2377 (1975)

158. R.W. Rozett & E.M. Peterssen, ibid., 47, 1301 (1975)

159. R.W. Rozett & E.M. Peterssen, ibid., 48, 817 (1976)

160. J.B. Justice Jr & T.L. Isenhour, ibid., 47, 2286 (1975)

161. G.L. Ritter, S.R. Lowry, T.L. Isenhour & C.L. Wilkins, ibid., 48, 591 (1976)

162. D.R. Burgard, S.P. Perone & J.L. Wiebers, ibid., 49, 1444 (1977)

163. pp136-72 of reference 35

164. J. Schechter & P.C. Jurs, Appl. Spectrosc., 27, 30 & 225 (1973)

165. P.C. Jurs in "Computer Representation and Manipulation of Chemical Information", eds. W.T. Wipke, S.R. Heller, R.J. Feldmann & E. Hyde, p265 (J. Wiley & Sons, New York, 1974)

166. G.S. Zander & P.C. Jurs, Anal. Chem., 47, 1562 (1975)

167. P. Edman, Acta Chem. Scand., 4, 283 (1950)

168. P. Edman & A. Begg, Eur. J. Biochem., 1, 80 (1967)

169. K. Biemann, Chimia, 14, 393 (1960)

170. P.J. Arpino & F.W. McLafferty in "Determination of Organic Structures by Physical Methods", Vol. 6, eds. F.C. Nachod, J.J. Zuckerman & E.W. Randall, pp1-89 (Academic Press, New York, 1976)

171. B. Sheldrick, Q. Rev. Chem. Soc., 24, 454 (1970)

172. H. Budzikiewicz, C. Djerassi & D.H. Williams, "Structure Elucidation of Natural Products by Mass Spectrometry", Vol. 2, p183 (Holden-Day, San Francisco, California, 1964)

173. M. Barber, P. Powers, M.J. Wallington & W.A. Wolstenholme, Nature (London), 212, 784 (1966)

174. K. Biemann, C. Cone, B.R. Webster & G.P. Arsenault, J. Am. Chem. Soc., 88, 5598 (1966)

175. M. Senn, R. Venkataraghavan & F.W. McLafferty, ibid., 88, 5593 (1966)

176. H.-K. Wipf, P. Irving, M. McCamish, R. Venkataraghavan & F.W. McLafferty, ibid., 95, 3369 (1973)

177. H. Nau, Angew. Chem. Int. Ed. Engl., 15, 75 (1976)

178. H. Nau & K. Biemann, Anal. Biochem., 73, 139, 154 & 175 (1976)

179. H. Lindley, Biochem. J., 126, 683 (1972)

180. M.O. Dayhoff & R.V. Eck, Comput. Biol. Med., 1, 5 (1970)

181. F. Sanger & E.O.P. Thompson, Biochem. J., 53, 353 (1953)

182. F. Sanger & H. Tuppy, ibid., 49, 481 (1951)

183. C.L. Weise & D.M. Desiderio, Comput. Biol. Med., 3, 437 (1973)

184. J.L. Wiebers, Anal. Biochem., 51, 542 (1973)

185. G.M. Pesyna & F.W. McLafferty in reference 170, pp91-155

186. R. Venkataraghavan, F.W. McLafferty & G.E. Van Lear, Org. Mass Spectrom., 2, 1 (1969)

187. F.W. McLafferty, K. -S. Kwok & G.M. Pesyna, J. Am. Chem. Soc., 95, 4185 (1973)

188. H.E. Dayringer, G.M. Pesyna, R. Venkataraghavan & F.W. McLafferty, Org. Mass Spectrom., 11, 529 (1976)

189. H.E. Dayringer & F.W. McLafferty, ibid., 11, 543 (1976)

190. H.E. Dayringer, F.W. McLafferty & R. Venkataraghavan, ibid., 11, 895 (1976)

191. H.E. Dayringer & F.W. McLafferty, ibid., 12, 53 (1977)

192. F.W. McLafferty, "Interpretation of Mass Spectra", 2nd ed. (W.A. Benjamin, Reading, Mass., 1973)

193. (a) E.G. Smith, "The Wiswesser Line Formula Chemical Notation", (McGraw-Hill, New York, 1968), (b) S.R. Heller & D.A. Koniver, J. Chem. Doc., 12, 55 (1972)

194. F.W. McLafferty, Anal. Chem., 49, 1441 (1977)

195. S.R. Lowry, T.L. Isenhour, J.B. Justice Jr, F.W. McLafferty, H.E. Dayringer & R. Venkataraghavan, ibid., 49, 1720 (1977)

196. F.W. McLafferty, R.H. Hertel & R.D. Villwock, Org. Mass Spectrom., 9, 690 (1974)

197. G.M. Pesyna, R. Venkataraghavan, H.E. Dayringer & F.W. McLafferty, Anal. Chem., 48, 1362 (1976)

198. G.M. Pesyna, F.W. McLafferty, R. Venkataraghavan & H.E. Dayringer, ibid., 47, 1161 (1975)

199. J. Lederberg, G.L. Sutherland, B.G. Buchanan, E.A. Feigenbaum, A.V. Robertson, A.M. Duffield & C. Djerassi, J. Am. Chem. Soc., 91, 2973 (1969)

200. A.M. Duffield, A.V. Robertson, C. Djerassi, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum & J. Lederberg, ibid., 91, 2977 (1969)

201. G. Schroll, A.M. Duffield, C. Djerassi, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum & J. Lederberg, ibid., 91, 7440 (1969)

202. Y. M. Sheikh, A. Buchs, A.B. Delfino, G. Schroll, A.M. Duffield, C. Djerassi, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum & J. Lederberg, Org. Mass Spectrom., 4, Supp., 493 (1970)

203. A. Buchs, A.B. Delfino, A.M. Duffield, C. Djerassi, B.G. Buchanan, E.A. Feigenbaum & J. Lederberg, Helv. Chim. Acta, 53, 1394 (1970)

204. A. Buchs, A.B. Delfino, C. Djerassi, A.M. Duffield, B.G. Buchanan, E.A. Figenbaum, J. Lederberg, G. Schroll & G.L. Sutherland, Adv. Mass Spectrom., 5, 314 (1971)

205. D.H. Smith, B.G. Buchanan, R.S. Engelmore, A.M. Duffield, A. Yeo, E.A. Feigenbaum, J. Lederberg & C. Djerassi, J. Am. Chem. Soc., 94, 5962 (1972)

206. D.H. Smith, B.G. Buchanan, R.S. Engelmore, H. Aldercreutz & C. Djerassi, ibid., 95, 6078 (1973)

207. D.H. Smith, B.G. Buchanan, W.C. White, E.A. Feigenbaum, C. Djerassi & J. Lederberg, Tetrahedron, 29, 3117 (1973)

208. L.M. Masinter, N.S. Sridharan, J. Lederberg & D.H. Smith, J. Am. Chem. Soc., 96, 7702 (1974)

209. L.M. Masinter, N.S. Sridharan, R.E. Carhart & D.H. Smith, ibid., 96, 7714 (1974)

210. R.G. Dromey, B.G. Buchanan, D.H. Smith, J. Lederberg & C. Djerassi, J. Org. Chem., 40, 770 (1975)

211. D.H. Smith, Anal. Chem., 47, 1176 (1975)

212. R.E. Carhart, D.H. Smith, H. Brown & N.S. Sridharan, J. Chem. Inf. Comput. Sci., 15, 124 (1975)

213. R.E. Carhart, D.H. Smith, H. Brown & C. Djerassi, J. Am. Chem. Soc., 97, 5755 (1975)

214. D.H. Smith, J.P. Konopelski & C. Djerassi, Org. Mass Spectrom., 11, 86 (1976)

215. C.J. Cheer, D.H. Smith, C. Djerassi, B. Tursch, J.C. Braekman & D. Daloze, Tetrahedron, 32, 1807 (1976)

216. B.G. Buchanan, D.H. Smith, W.C. White, R.J. Gritter, E.A. Feigenbaum, J. Lederberg & C. Djerassi, J. Am. Chem. Soc., 98, 6168 (1976)

217. C. Lageot, J. Organomet. Chem., 96, 355 (1975)

218. C. Lageot, ibid., 102, C7-C8 (1975)

219 C. Lageot, Org. Mass Spectrom., 11, 1194 (1976)

220. A.B. Delfino & A. Buchs, Helv. Chim. Acta, 55, 2017 (1972)

221. F.W. McLafferty, "Mass Spectral Correlations", (American Chemical Society, Washington, D.C., 1963)

222. H. Budzikiewicz, C. Djerassi & D.H. Williams, "Mass Spectrometry of Organic Compounds", (Holden-Day, San Francisco, 1967)

223. B.R. Kowalski & C.F. Bender, J. Am. Chem. Soc., 94, 5632 (1972)

224. B.R. Kowalski, Comput. Chem. Biochem. Res., 2, 2 (1974)

225. D.L. Duewer, B.R. Kowalski & T.F. Schatzki, Anal. Chem., 47, 1573 (1975)

226. D.L. Duewer & B.R. Kowalski, ibid., 47, 526 (1975)

227. P.J. Simon, B.C. Giessen & T.R. Copeland, ibid., 49, 2285 (1977)

228. N.A.B. Gray, ibid., 48, 2265 (1976)

229. D.L. Duewer & H. Freiser, ibid., 49, 1940 (1977)

230. S. Hanessian, Methods Biochem. Anal., 19, 105 (1971)

231. J.A. McCloskey in "Basic Principles of Nucleic Acid Chemistry", Vol. 1, ed. P.O.P.Ts'o, p209 (J. Wiley & Sons, New York, 1974)

232. D.C. DeJongh in "Synthetic Procedures in Nucleic Acid Chemistry, Vol.2. Physical and Physicochemical Aids in Characterisation and in Determination of Structure", eds. W.W. Zorbach & R.S. Tipson, p145 (J. Wiley & Sons, New York, 1973)

233. C. Hignite in "Biochemical Applications of Mass Spectrometry", ed. G.R. Waller, p429 (J. Wiley & Sons, New York, 1972)

234. D.C. DeJongh, T. Radford, J.D. Hribar, S. Hanessian, M. Bieber, G. Dawson & C.C. Sweeley, J. Am. Chem. Soc., 91, 1728 (1969)

235. F.H. Field, Acc. Chem. Res., 1, 42 (1968)

236. J. Block, Adv. Mass Spectrom., 4, 791 (1968)

237. J.H. McReynolds, N.W. Flynn, R.R. Sperry, D. Fraisse & M. Anbar, Anal. Chem., 49, 2121 (1977)

238. (a) H.-R. Schulten & H.D. Beckey, Org. Mass Spectrom., 7, 861 (1973), (b) W.D. Lehmann, H.-R.Schulten & H.D. Beckey, ibid., 7, 1103 (1973), (c) J. Moor & E.S. Waight, ibid., 9, 903 (1974), (d) H.-R. Schulten, Methods Biochem. Anal., 24, 313 (1977)

239. (a) J.A. McCloskey, J.H. Futrell, T.A. Elwood, K.H. Schram, R.P. Panzica & L.B. Townsend, J. Am. Chem. Soc., 95, 5762 (1973), (b) M.S. Wilson & J.A. McCloskey, ibid., 97, 3436 (1975), (c) P. Brown, G.R. Pettit & R.K. Robins, Org. Mass Spectrom., 2, 521 (1969), (d) J.R. Majer & A.A. Boulton, Methods Biochem. Anal., 21, 467 (1973)

240. S.J. Shaw, D.M. Desiderio, K. Tsuboyama & J.A. McCloskey, J. Am. Chem. Soc., 92, 2510 (1970)

241. M.J. Robins, S.R. Naik & A.S.K. Lee, J. Org. Chem., 39, 1891 (1974)

242. K. Biemann & J.A. McCloskey, J. Am. Chem. Soc., 84, 2005 (1962)

243. H.A. Howlett, M.W. Johnson, A.R. Trim, J. Eagles & R. Self, Anal. Biochem., 39, 429 (1971)

244. S. Hanessian, D.C. DeJongh & J.A. McCloskey, Biochim. Biophys. Acta, 117, 480 (1966)

245. M.J. Robins & G.L. Basom, Can. J. Chem., 51, 3161 (1973)

246. pp585-9 of reference 222

247. J.M. Rice, G.O. Dudek & M. Barber, J. Am. Chem. Soc., 87, 4569 (1965)

248. T. Nishiwaki, Tetrahedron, 22, 3117 (1966) & 23, 1153 (1967)

249. S.M. Hecht, A.S. Gupta & N.J. Leonard, Biochim. Biophys. Acta, 182, 444 (1969)

250. E. White V, P.M. Krueger & J.A. McCloskey, J. Org. Chem., 37, 430 (1972)

251. pp592-4 of reference 222

252 J.H. Lister, "Chemistry of Heterocyclic Compounds. Vol. 24. Fused Pyrimidines, Part 2: Purines", pp518-23 (J. Wiley & Sons, New York, 1971)

253. D.S. Letham, J.S. Shannon & I.R. McDonald, Proc. Chem. Soc. London, 1964, 230

254. J.S. Shannon & D.S. Letham, N.Z.J. Sci., 9, 833 (1966)

255. J.M. Rice & G.O. Dudek, J. Am. Chem. Soc., 89, 2719 (1967)

256. N.J. Leonard, S.M. Hecht, F. Skoog & R.Y. Schmitz, Proc. Nat. Acad. Sci. U.S.A., 59, 15 (1968)

257. J. Heiss, K.-P. Zeller & W. Voelter, Org. Mass Spectrom., 3, 181 (1970)

258. S.M. Hecht, Anal. Biochem., 44, 262 (1971)

259. S.M. Hecht & J.J. McDonald, ibid., 47, 157 (1972)

260. P.B. Farmer, A.B. Foster, M. Jarman & M.J. Tisdale, Biochem. J., 135, 203 (1973)

261. S.M. Hecht, A.S. Gupta & N.J. Leonard, Anal. Biochem., 30, 249 (1969)

262. J.M. Rice & G.O. Dudek, Biochem. Biophys. Res. Commun., 35, 383 (1969)

263. L.B. Townsend & R.K. Robins, J. Heterocycl. Chem., 6, 459 (1969)

264. P.F. Crain, J.A. McCloskey, A.F. Lewis, K. Schram & L.B. Townsend, ibid., 10, 843 (1973)

265. M.J. Robins, J.R. McCarthy Jr, R.A. Jones & R. Mengel, Can. J. Chem., 51, 1313 (1973)

266. K. Biemann & W. McMurray, Tetrahedron Lett., 1965, 647

267. J.G. Liehr, D.L. von Minden, S.E. Hattox & J.A. McCloskey, Biomed. Mass Spectrom., 1, 281 (1974)

268. M.J. Robins, R.H. Hall & R. Thedford, Biochemistry, 6, 1837 (1967)

269. p15 of reference 35

270. "Chemometrics: Theory and Application", ed. B.R. Kowalski (American Chemical Society, Washington, D.C., 1977)

271. A.M. Harper, D.L. Duewer, B.R. Kowalski & J.L. Fasching in reference 270, pp14-52

272. P.C. Jurs, Anal. Chem., 43, 22 (1971)

273. (a) H. Rotter & K. Varmuza, Org. Mass Spectrom., 10, 874 (1975), (b) K. Varmuza & H. Rotter, Monatsh. Chem., 107, 547 (1976)

274. N.M. Abramson, "Information Theory and Coding", pp11-44 (McGraw-Hill, New York, 1963)

275. "Biomedical Computer Programs", ed. W.J. Dixon, pp233-254 (University of California Press, Berkeley & Los Angeles, California & London, 1973)

276. (a) T.W. Anderson, "An Introduction to Multivariate Statistical Analysis" (J. Wiley & Sons, New York, 1958), (b) C.R. Rao, "Linear Statistical Inference and Its Applications", pp487-493 (J. Wiley & Sons, New York, 1965)

277. R.J. Harris, "A Primer of Multivariate Statistics" (Academic Press, New York, 1975)

278. pp101-113 of reference 277

279. pp231-233 of reference 277

280. C.R. Rao, "Advanced Statistical Methods in Biometric Research", pp258-271 (J. Wiley & Sons, New York, 1952)

281. L.B. Sybrandt & S.P. Perone, Anal. Chem., 43, 382 (1971)

282. P.C. Jurs, B.R. Kowalski, T.L. Isenhour & C.N. Reilley, Anal. Chem., 41, 690 (1969) & 42, 1387 (1970)

283. pp13-16 & pp173-179 of reference 35

284. S.M. Hecht, A.S. Gupta & N.J. Leonard, Anal. Biochem., 38, 230 (1970)

285. M.J. Robins & E.M. Tripp, Biochemistry, 12, 2179 (1973)

286. W.J. Burrows, F. Skoog & N.J. Leonard, ibid., 10, 2189 (1971)

287. M.J. Robins, Y. Fouron & R. Mengel, J. Org. Chem., 39, 1564 (1974)

288. R.J. Suhadolnik, B.M. Chassy & G.R. Waller, Biochim. Biophys. Acta, 179, 258 (1969)

289. J.M.J. Tronchet & R. Graf, Helv. Chim. Acta, 56, 2689 (1973)

290. S.H. Eggers, S.I. Biedron & A.O. Hawtrey, Tetrahedron Lett., 1966, 3271

291. R.H. Hall, L. Csonka, H. David & B. McLennan, Science, 156, 69 (1967)

292. M.W. Logue & N.J. Leonard, J. Am. Chem. Soc., 94, 2842 (1972)

293. J.M.J. Tronchet & M.F. Perret, Helv. Chim. Acta, 55, 2121 (1972)

294. N. Mariaggi & R. Teoule, Bull. Soc. Chim. Fr., 1976, 1595

295. S.M. Hecht, L.H. Kirkegaard & R.M. Bock, Proc. Nat. Acad. Sci. U.S.A., 68, 48 (1971)

296. J. Ulrich, J. Cadet & R. Teoule, Org. Mass Spectrom., 7, 543 (1973)

297. J.J. Dolhun & J.L. Wiebers, ibid., 3, 669 (1970)

298. M.J. Robins, R. Mengel, R.A. Jones & Y. Fouron, J. Am. Chem. Soc., 98, 8204 (1976)

299. U. Sequin & Ch. Tamm, Helv. Chim. Acta, 55, 1196 (1972)

300. W.J. Burrows, D.J. Armstrong, F. Skoog, S.M. Hecht, J.T.A. Boyle, N.J. Leonard & J. Occolowitz, Science, 161, 691 (1968)

301. W.J. Burrows, D.J. Armstrong, F. Skoog, S.M. Hecht, J.T.A. Boyle, N.J. Leonard & J. Occolowitz, Biochemistry, 8, 3071 (1969)

302. D.J. Armstrong, P.K. Evans, W.J. Burrows, F. Skoog, J.-F. Petit, J.L. Dahl, T. Steward, J.L. Strominger, N.J. Leonard, S.M. Hecht & J. Occolowitz, J. Biol. Chem., 245, 2922 (1970)

303. S.M. Hecht, N.J. Leonard, J. Occolowitz, W.J. Burrows, D.J. Armstrong, F. Skoog, R.M. Bock, I. Gillam, & G.M. Tener, Biochem. Biophys. Res. Commun., 35, 205 (1969)

304. S.M. Hecht, N.J. Leonard, R.Y. Schmitz, & F. Skoog, Phytochemistry, 9, 1173 (1970)

305. H. Kasai, Z. Ohashi, F. Harada, S. Nishimura, N.J. Oppenheimer, P.F. Crain, J.G. Liehr, D.L. von Minden & J.A. McCloskey, Biochemistry, 14, 4198 (1975)

306. K.L. Nagpal & J.P. Horwitz, J. Org. Chem., 36, 3743 (1971)

307. W.C. Butts, J. Chromatogr. Sci., 8, 474 (1970)

308. M.J. Robins, M. MacCross & A.S.K. Lee, Biochem. Biophys. Res. Commun., 70, 356 (1976)

309. G.G. Deleuze, J.D. McChesney & J.E. Fox, ibid., 48, 1426 (1972)

310. M.P. Schweizer, K. McGrath & L. Baczynskyj, ibid., 40, 1046 (1970)

311. Z. Ohashi, K. Murao, T. Yahagi, D.L. von Minden, J.A. McCloskey & S. Nishimura, Biochim. Biophys. Acta, 262, 209 (1972)

312. S.M. Hecht, N.J. Leonard, W.J. Burrows, D.J. Armstrong, F. Skoog & J.Occolowitz, Science, 166, 1272 (1969)

313. W.J. Burrows, D.J. Armstrong, M. Kaminek, F. Skoog, R.M. Bock, S.M. Hecht, L.G. Dammann, N.J. Leonard & J. Occolowitz, Biochemistry, 9, 1867 (1970)

314. T. Huynh-Dinh, A. Kolb, C. Gouyette, J. Igolen & S. Tran-Dinh, J. Org. Chem., 40, 2825 (1975)

315. H.S. El Khadem & El S.H. El Ashry, Carbohydr. Res., 32, 339 (1974)

316. G. Giovanninetti, L. Nobile, A. Andreani, A. Ferranti, M. Amorosa & J. Defaye, ibid., 27, 243 (1973)

317. G. Giovanninetti, L. Nobile, M. Amorosa & J. Defaye, ibid., 21, 320 (1972)

318. L. Nobile, G. Giovanninetti, T.P. Balbi, M. Amorosa & J. Defaye, ibid., 24, 489 (1972)

319. J. Defaye & Z. Machon, ibid., 24, 235 (1972)

320. This work, 1979

321. J.T. Clerc, P. Naegeli & J. Seibl, Chimia, 27, 639 (1973)

322. K.-L.H. Ting, R.C.T. Lee, G.W.A. Milne, M. Shapiro & A.M. Guarino, Science, 180, 417 (1973)